Coarse Memory and Plausible Narratives*

Francesco Bilotta Giacomo Manferdini

June 27, 2024

Abstract

In a political economy framework, we study how false narratives emerge in response to imperfections in recipients' memory. Coarse memory allows voters to retrieve marginal frequencies of past policies and outcomes, but prevents them from understanding their correlations. Politicians exploit such vagueness by designing narratives inflating the effectiveness of their preferred policies. We find that plausible narratives can be less optimistic about more implemented policies. In a probabilistic voting model we show that opposing narratives are polarized and that in the long run political cycles arise. Our mechanism is consistent with an analysis of U.S. congress members' rhetorical strategies.

Keywords: Narratives, Memory, Communication, Probabilistic Voting

*Francesco Bilotta: Bocconi University, Department of Economics; francesco.bilotta2@phd.unibocconi.it. Giacomo Manferdini: Bocconi University, Department of Economics; giacomo.manferdini@phd.unibocconi.it. We are especially grateful to our advisors Nicola Pavoni and Massimo Morelli for constant guidance and encouragement. We also thank Pierpaolo Battigalli, Luca Bellodi, Alberto Binetti, Luca Braghieri, Andrea Gallice, Nicola Gennaioli, Charles Manski, Nina Nikiforova, Salvo Nunnari, Pietro Ortoleva, Burkhard Schipper and Guido Tabellini as well as participants at Munich Workshop on Beliefs Narratives and Memory, Bocconi Theory Brown Bag, 17th UniTo-CCA PhD Workshop, 1st Bocconi-CCA-Cornell Workshop in Political Economy, UC Davis Microeconomic Theory Seminar for valuable comments; Hugo Lavenant and Lucas Pesenti for mathematical advice. All errors remain our own.

> Even if the true scientists should all recognize the limitations of what they can do in the field of human affairs, so long as the *public expects more* there will always be *some who will pretend*, and perhaps honestly believe, that they can do more to meet popular demands than is really in their power. It is often difficult enough for the expert, and certainly in many instances *impossible* for the layman, *to distinguish* between legitimate and illegitimate claims advanced in the name of science.
>
> Hayek (1974), Nobel Prize Lecture

As emphasized in Hayek (1974), economies are complex systems involving variables whose relationships are hard to observe and measure. As a consequence, the task of distinguishing objectively valid claims about their behavior from plausible – yet fallacious – ones is hard, especially for the general public. A natural reaction to this state of ignorance is a demand for reassuring explanations, which may be met by individuals who pretend to be capable of better interventions than they truly are. This view is supported by evidence about the recent evolution of the political debate, which reveals how candidates increasingly rely on crafting a preferential communication channel with their electorate, in order to persuasively share their own worldviews. For instance, take the rise of spin doctors and political consultants (Sheingate, 2016) or the increase in political advertisements expenditure in the US (The Guardian, 2022). In this paper, we formalize the idea that the supply of false narratives emerges in response to a demand for plausible and promising explanations by a boundedly rational electorate, which struggles to perceive and recall correlations between economic variables of interest. Then, we study its implications for political competition, in a stylized model of retrospective voting.

Specifically, we consider an economy where outcome is produced by the implementation of policies, in a potentially stochastic fashion, as dictated by an objectively true model of the economy. Voters have a coarse memory of the environment: they correctly recall the frequency with which different policy interventions have been implemented and the frequency with which different economic outcomes have been attained, but they completely ignore which policies led to which outcomes. Hence, they cannot reconstruct the true model. While our assumption appears stark, we think it approximates well the complexity of understanding economic policies, whose implementation is typically staggered and whose effects are often delayed and subject to noise. Moreover, it aligns well with laboratory evidence demonstrating how agents may fail to correctly understand correlations in the data they observe (Ambuehl and Thysen, 2024; Eyster and Weizsacker, 2016), particularly in noisy environments (Fréchette et al., 2024), as well as with psychological studies indicating that people struggle to accurately remember associations (Kahana, 2012) and often misattribute events to incorrect sources (Schacter, 1999). In this context, we picture narratives as devices that offer a description of the link between policies and outcomes, adopting a definition which is close to that of if-then clauses in the Bayesian psychology of verbal reasoning and argumentation (Oaksford and Chater, 2020, for a review). Formally, we consider stochastic maps associating to each policy a probability distribution over outcome levels, describing its claimed potential impact. Our main observation is that, beyond the true model of the economy, many different maps of this kind are plausible, in the sense that they generate the outcome distribution recalled by the voter given that of policy interventions, rationalizing voters' coarse memory. As a consequence, coarseness offers an opportunity to strategic politicians, who can spread plausible but typically false narratives, inflating the effectiveness of their preferred policy against the others.

In our setting, political competition resembles a blame game between partners (the politicians) who jointly produced a common good of ambiguous quality (a history of outcomes) and must defend their interventions before an external arbiter (the voter). Given the voter's coarse memory, each politician has limited "property rights" over the outcomes their interventions produced. Consequently, both competitors strive to claim credit for positive outcomes in the observed history while discrediting their opponent. They structure their narratives to attribute the best outcomes to their own interventions and the worst ones to their opponents. In this sense, narratives serve as merit-stealing and buck-passing devices, allowing the proponents of inferior interventions to free ride on the positive outcomes generated by superior ones. However, voters' memory imposes some discipline on this manipulation, linking the plausibility of narratives with the frequency of policy implementation. Specifically, we find that the more frequently a policy has been implemented in the past, the harder it is to claim a false distribution for its effects while satisfying the plausibility constraint. Thus, plausible narratives can be more optimistic about the effectiveness of a policy when it has been implemented less frequently in the past, whereas they must be closer to the truth for policies that have been implemented more often. This is the main mechanism underlying our model and we find it is compatible with the rhetorical strategies adopted by U.S. congress members' when debating over the effectiveness of the Affordable Care Act on Twitter.

In turn, the mechanism has numerous implications for political competition, also compatible with anecdotal evidence. For starters, our model produces an implicit cost of ruling: the set of narratives available to the incumbent shrinks around the true model during their mandate, because of the more frequent implementation of their preferred policy, while the opposite happens for the competitor. In this sense, narratives may be a source of disadvantage for the incumbent, contributing to explain the established finding that, on average, a government steadily loses votes between election periods[1]. Relatedly, our model suggests how narratives may be a force fostering instability and cycles in political power[2]. When the frequency of policy implementation evolves dynamically as the result of elections where vot-

[1] See, for instance, (Nannestad and Paldam, 2003; Paldam, 1986)

[2] We emphasize that our goal is not to match historical data regarding the recurrence of policies, which would require a more complex model accounting for numerous other factors. Instead, we aim to make the qualitative point that narratives may be one of these factors.

ers are captured by the most promising narrative, the economy is eventually trapped in a state where plausibility has the least bite and, ultimately, where parties govern with the same frequency in the long run. Hence, narratives erode objective differences between policies, depressing the frequency of implementation of the best one. Moreover, our model predicts heterogeneity in the structure of narratives. In particular, it predicts that politicians who bet on unconventional policies may support them with potentially sensationalistic claims, while those who defend more seasoned ones often have to take some blame, and will focus on claiming incompetence for their opponents' proposals[3]. Finally, we find that, since opponents aim at defending different policies, narratives advanced by opposing candidates will naturally be polarized, in the sense of disagreeing about the effectiveness of the same policy intervention as much as plausibility allows[4].

The rest of the paper is organized as follows. Section 1 illustrates a binary example which allows us to make the above intuitions precise and supports them with some suggestive evidence. The general problem of designing the best plausible narrative to defend an intervention is treated in Section 2, where connections with the theory of Optimal Transport and Partial Identification are drawn. In Section 3 we study the implications of the problem in a static game of political competition, while in Section 4 we analyze the dynamics arising from myopic repetitions of this game. Finally, Section 5 takes stocks. We conclude this Section discussing how our paper relates to the existing literature.

Related Literature Our paper contributes mainly to the literature on narratives in economics (Shiller, 2017), see Barron and Fries (2024) for a recent review. Within it, one strand of theoretical work (Eliaz and Spiegler, 2020; Eliaz et al., 2023; Horz and Kocak, 2022) describes narratives as directed acyclic graphs (Pearl, 2009) building on the seminal contribution of Spiegler (2016), who showed how they can be used to model misperceptions in the directions of causal relationships between economic variables (Spiegler, 2020, for a rewiew). In this approach, when an agent believes that his environment is described by a DAG R, potentially different from the true one R^*, he imposes the conditional independence relation implied by R on the true data generating process p. In this way, he may interpret spurious correlations between variables causally and hold biased beliefs about the effects of interventions. We differ from this approach since, in our setting, the direction of causality is clear – policies are direct

[3] This rhetoric opportunity seems to be exploited by entrant (populist) parties in choosing their flagship policies, sometimes resulting in considerable success. Take, for instance, the rise of the Five-Star Movement in Italy, which proposed a universal basic income (never experimented in the Italian economy) and claimed that it would have revamped employment. As another case, consider Matteo Salvini's rebranding of Lega Nord in Italy, based on a flat taxation system presented as a solution for the Italian economy.

[4] Disagreement between politicians about the effectiveness of policies appears an ordinary features of the political debate. As an example, consider the diametrically opposing narratives regarding the impact of immigration on the U.S. economy during the 2016 presidential election. Trump: *"Decades of record immigration have produced lower wages and higher unemployment for our citizens, especially for African-American and Latino workers."* Clinton: *"Comprehensive immigration reform will grow our economy and keep families together – and it is the right thing to do."*

causes of outcomes – but narratives inform beliefs about causal strength, claiming potentially over- or under-stated effectiveness for each intervention. Hence, while this strand of literature typically focuses on easy fix narratives based on reverse causality, we focus on alternative causes, competing for attribution of the best outcomes via buck-passing and merit-stealing. In metaphorical terms, while Eliaz and Spiegler (2020) is an "extensive margin" theory of narratives, ours is an "intensive margin" one. Beyond these important conceptual differences, treating narratives as stochastic maps allows us to solve an explicit narrative design problem instead of relying on a long run equilibrium concept.

We differ more markedly from a second strand of the literature (Aina, 2021; Ispano, 2023; Izzo et al., 2021; Schwartzstein and Sunderam, 2021) which builds on Bayesian persuasion (Kamenica and Gentzkow, 2011), modeling narratives as alternative information structures offered to an agent, who is taken to update beliefs via the model that gives the highest likelihood to the observed signals, capturing a form of abductive reasoning. Importantly, this implies that persuasion by false narratives becomes increasingly difficult as sample size increases[5]. Instead, coarse memory prevents the agent from ever point identifying the true model: on the contrary, in the long run, the economy is trapped in a state where identification is the hardest. In particular, Ispano (2023) builds on Aina (2021) to characterize, in a binary example, the posterior beliefs which can be induced providing an agent with a signal structure compatible with a given prior and an observed distribution of signal realizations. Their construction, like ours, leverages the law of total probabilities, and hence also results in a linear constraint. Nonetheless, despite the graphical similarity, the two constructions are carried out in different spaces (posterior vs. conditional beliefs). Indeed, in our framework, agents perform no Bayesian updating.

Compared with the two previous approaches, a novel aspect of our proposal is, in our opinion, to link the diffusion of false narratives with a specific limitation in voters' cognition, namely their inability to detect and recall correlations, especially in complex environments. This emerges in numerous experimental papers related to agents' perception of correlations. For instance, Ambuehl and Thysen (2024) find that about half of their subjects fails to use any form of correlational information available in the data. Similarly, Eyster and Weizsacker (2016) estimate that only a small fraction of their experimental units are able to properly appreciate the covariance structure of asset portfolios. Fréchette et al. (2024), in recent work in progress, re-iterate the finding, showing that the effect is exacerbated in noisy environments. Overall, we take these findings to support our view that agents may be substantially uncertain about correlation structures and willing to accept plausible ones. We stress how this is different from

[5] As noted also by Barron and Fries (2023), in the large sample limit, the true model will almost surely maximize the likelihood of the evidence. Hence, the only possibility to have persuasion in this case is to force that the receiver is never exposed to the true model (nor that she can back it up from data, which she perceives perfectly in their setting).

correlation neglect, which is the agent's tendency to treat potentially correlated signals as independent, leading to overprecision in Bayesian updating and extremism in agents' beliefs (Ortoleva and Snowberg, 2015, for instance).

Our focus on correlations leads to coarse memory, which differs from models founding memory on databases of observations and introducing bias in the recall process as the effect of context or time (Bordalo et al., 2020; Fudenberg et al., 2022, for instance). Instead, we are closer to the methodological work of Battigalli and Generoso (2021) who propose the first representation of sequential games explicitly separating the description of players' memory from that of the rules of the game. Among their examples of memory correspondences, statistical memory is close to our notion of coarseness.

We contribute to the literature on dynamic political competition, showing how plausible narratives allow riding on the rare implementation of a policy to inflate its value, driving policy cycles. In this sense, we differ from Levy et al. (2022) where policy cycles arise since a part of the polity holds a simplified worldview leading to preferences for few extreme interventions which are more intense than those of the correctly specified faction. At the same time, we show how coarse memory, preventing voters from debunking many false narratives, is an important limit to the accountability of politicians: we thus add to models considering how retrospective voting may be subject to bias (Esponda and Pouzo, 2017, for instance, consider the effects of sample selection).

Finally, we see a contribution in importing the conceptual lessons of partial identification (Manski, 1995) in behavioral economics. In particular, we relate to papers studying identification from contaminated and corrupted data (Horowitz and Manski, 1995) and the dynamics of identification bounds (Manski, 2004). Similarly, we connect to the theory of optimal transport (Galichon, 2018, for an excellent introduction) expanding the range of its applications to behavioral economics problems.

1 Simple Example and Suggestive Evidence

We picture the economy as a system transforming policy inputs into economic outcomes, where the impact of the former on the latter is probabilistic. To fix ideas, consider two policies, *h*igh and *l*ow and two possible outcomes, *g*ood and *b*ad. For each policy $a \in \{h, l\}$ we posit an *interventional distribution* over outcome levels[6] $\mu^*(a) \in \Delta(\{g, b\})$, which describes the *objective* odds of seeing the good outcome when intervention a is implemented. Since outcome is binary, $\mu^*(a)$ can be identified with the probability of the good outcome under intervention a, $\mu^*_a \in [0, 1]$, which can be thought of as the *effectiveness* of policy a. In the following, we refer

[6] Notation: we denote by $\Delta(X)$ the set of probability distributions over the set X.

to $\mu^* \equiv (\mu_h^*, \mu_l^*) \in [0,1]^2$ as the *true model* of the economy[7]. To substantiate our naming, we let the high policy be more effective than the low[8]: $\mu_h^* > \mu_l^*$.

We consider a representative voter (she) who *completely ignores* μ^*, hence the best policy, and has only *coarse memory* of the economy's input-output performance: she fails to record which interventions happened with which outcomes[9] but correctly recalls the frequency of implementation for policy interventions and the frequency of realization for outcome levels. In our binary example, coarse memory is described by two variables: $\alpha_h \in (0,1)$, the frequency of implementation of the high policy h; and $\nu_g \in (0,1)$, the frequency of g, the good outcome. Since α_h and ν_g are objective quantities, the true model μ^* links the two variables at any observation time. In particular, by the law of total probabilities, $\nu_g \in (0,1)$ is the average of the vector μ^* weighted by α_h,

$$\nu_g(\alpha_h, \mu^*) = \alpha_h \mu_h^* + (1-\alpha_h)\mu_l^*. \tag{1}$$

A narrative consists of an alternative pair of interventional distributions $(\mu(a))_{a\in\{h,l\}}$ which we identify with the effectiveness claimed for the two policies $\mu \equiv (\mu_h, \mu_l) \in [0,1]^2$. Our main observation is that coarse memory, despite its limitations, still allows the voter to perform a simple test on the narratives she is proposed. We can formalize this aspect by positing that the agent performs the following *coarse falsification procedure*. When the voter recalls frequencies $(\alpha_h, \nu_g) \in [0,1]^2$, she:

1. computes the frequency of good outcome implied, through the law of total probabilities, by narrative μ and policy frequency α_h, that is $\alpha_h\mu_h + (1-\alpha_h)\mu_l$;
2. retrieves from memory the frequency of good outcome, ν_g;
3. considers narrative $\mu = (\mu_h, \mu_l)$ *plausible* if and only if the two match, namely

$$\alpha_h\mu_h + (1-\alpha_h)\mu_l = \nu_g. \tag{2}$$

Hence, despite coarseness, the voter is still able to falsify some deceptive claims[10]. Note that, by Equation 1, the true model μ^* always passes the plausibility test. At the same time, given

[7] We stress that this is a very stylized but objective description of how the economy works, and it is intended to summarize more fine-grained accounts, where the impact of policies on outcomes may be mediated by complex mechanisms, possibly involving chains of variables lying outside the direct control of the policymaker or exogenous shocks.

[8] Of course, this is without any loss of generality. Given two competing policies, there will always be one which is objectively more beneficial to a given economy (at least at a given moment of observation).

[9] In particular, the voter cannot point-identify μ^* from memory

[10] For instance, imagine that one of the policies proposes subsidy packages and is supported by a left-wing politician who insists on demand-driven motivations for growth to claim high effectiveness for it; on the other hand, the competing policy proposes tax cuts and is supported by a right-wing politician who insists on supply-driven reasoning. The intuition beyond coarse falsification is pretty straightforward: if, say, tax cuts have been implemented very frequently, but the outcome has rarely been good, the voter should not accept narratives that ascribe very high effectiveness to tax cuts.

a coarse memory (α_h, ν_g), many other narratives will satisfy the requirement of plausibility. They can be collected in the set of *plausible narratives*

$$\mathcal{M}(\alpha_h, \nu_g) = \left\{ (\mu_h, \mu_l) \in [0,1]^2 \;\middle|\; \alpha_h \mu_h + (1-\alpha_h)\mu_l = \nu_g \right\} \tag{3}$$

that, in a natural sense, are the models *behaviorally equivalent* to the true one (confront Equations 1 and 2) when restricting to the coarse memory (α_h, ν_g).

We consider a politician (he) who has no flexibility in choosing his platform but can only craft an effective narrative to defend his commitment policy[11]. Since the representative voter accepts a narrative only if it is plausible, the politician is constrained to the set $\mathcal{M}(\alpha_h, \nu_g)$. Although many objectives are possible, a natural one is to maximize the utility the voter anticipates from the implementation of the politician's commitment, normalizing the payoff from the good outcome to 1 and from the bad one to 0. Considering, without loss, the politician committed to the high policy – call him H – the problem is

$$\begin{aligned} \max_{(\mu_h^H, \mu_l^H) \in [0,1]^2} \quad & \mu_h^H \\ \text{subj. to:} \quad & \alpha_h \mu_h^H + (1-\alpha_h)\mu_l^H = \nu_g \end{aligned} \tag{P$_{\text{simple}}$}$$

and has the following immediate solution

$$\hat{\mu}_h^H = \min\left\{1, \frac{\nu_g}{\alpha_h}\right\} \qquad \hat{\mu}_l^H = \max\left\{0, \frac{\nu_g - \alpha_h}{1-\alpha_h}\right\}. \tag{4}$$

Equation 4 reflects simple economic reasoning: the optimal narrative for a politician involves stating that his commitment policy is fully effective whenever he can. Clearly, memories for which this claim is plausible are those with $\alpha_h \leq \nu_g$. In this case, the policy has been proposed rarely enough that the politician can plausibly state: "My policy has not been implemented frequently, but the few times it was tested, it worked!". To make this claim plausible, the politician complements it attributing the residual outcome to the opponent's policy. Instead, for memories with $\alpha_h > \nu_g$, this narrative does not satisfy the plausibility requirement. In this case, the policy has been implemented too frequently to claim it always worked. Hence, the politician falls back to a narrative that recognizes some flaws in his own policy, admitting its incomplete effectiveness. However, he does so optimally, claiming the opponent's policy is completely ineffective, so that he can still attribute all the good outcome to his own policy. In words, these are the situations when the politician can plausibly state: "My policy may not be perfect, but my opponent's measures are completely useless!". In-

[11] This allows us to focus on the direct effect of narratives on political competition. Moreover, in reality, commitment may arise for various reasons, such as ideological preferences or as a consequence of voters' decreasing trust in the political system (Bellodi et al., 2023).

deed, interpreting Equation 4 from an econometric standpoint, the optimal narrative consists in claiming the *upper* identification bound for the effectiveness of one's own commitment while claiming the *lower* identification bound for the effectiveness of the opponent's. In Theorem 1, we will show how to extend this logic of optimal rearrangement of outcome to the general problem.

We can gain insight into the solution of P_{simple} and its comparative statics by graphical reasoning. To this end, note that substituting Equation 1 in Equation 3 we obtain the following expression[12] for $\mathcal{M}$ in terms of the true model of the economy μ^*

$$\mathcal{M}(\alpha_h, \mu^*) = \left\{ (\mu_h, \mu_l) \in [0,1]^2 \middle| \mu_l - \mu_l^* = -\left(\frac{\alpha_h}{1-\alpha_h} \right) (\mu_h - \mu_h^*) \right\}. \tag{5}$$

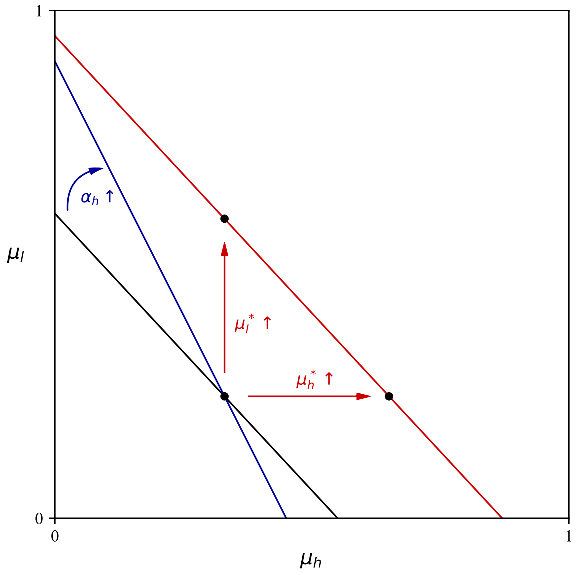
That is, $\mathcal{M}$ is the unique segment in $[0,1]^2$ that passes through the point $\mu^* = (\mu_h^*, \mu_l^*)$ and whose slope, letting μ_l be on the y-axis and μ_h on the x-axis, is $-\frac{\alpha_h}{1-\alpha_h}$, as depicted in Figure 1.Therefore, plausibility imposes a link on how a narrative may detach from the true model along the two policy dimensions: plausible narratives have to *substitute* the effectiveness of one policy versus the other. Indeed, if μ_h increases by one percentage point, μ_l has to decrease by $\frac{\alpha_h}{1-\alpha_h}$ percentage points for the narrative to remain plausible. That is, the relative frequency of implementation of one policy against the other, measured by α_h, regulates how "costly" it is to inflate the effectiveness of a policy versus the other. In particular, as α_h increases, the segment rotates clockwise around the point $\mu^* = (\mu_h^*, \mu_l^*)$ and the plausible values for the effectiveness of the policy h (which are the projection of the segment on the x-axis) shrink around the true one (the value μ_h^*) leaving less space for inflating its effectiveness. Of course, things work the other way around for policy l. This is the mechanism through which the value of P_{simple}, which is $\hat{\mu}_h^H$, decreases (weakly) in α_h, as it is evident from Equation 4. Additional insight on the solution may be obtained substituting Equation 1 in Equation 4 so that

$$\hat{\mu}_h^H = \min\left\{1, \mu_h^* + \left(\frac{1-\alpha_h}{\alpha_h}\right)\mu_l^*\right\} \quad \hat{\mu}_l^H = \max\left\{0, \mu_l^* - \left(\frac{\alpha_h}{1-\alpha_h}\right)(1-\mu_h^*)\right\}. \tag{6}$$

First, these expressions show how optimality leads the politician to depart from the true model in his announcement, overshooting the effectiveness of his own policy and understating the opponent's one. Indeed, the true effectiveness μ_h^* is a lower bound to the announcement $\hat{\mu}_h^H$ for any memory parameters: it is always feasible to just tell the truth. The converse holds for $\hat{\mu}_l^H$. In particular, note how, in the limit $\alpha_h \to 1$, the politician is forced to truth-telling, while he is increasingly less constrained as $\alpha_h \to 0$. Secondly, the expressions allow us to study the impact of a change in the true model μ^*. Unsurprisingly, when μ_h^* is higher,

[12] Of course, such an expression is accessible to the modeler, but not to the representative voter, who ignores μ^*. The distinction between the two perspectives is common to settings where some players have a limited conception of their environment, such as in the case of unawareness.

Figure 1: Plausible narratives and comparative statics

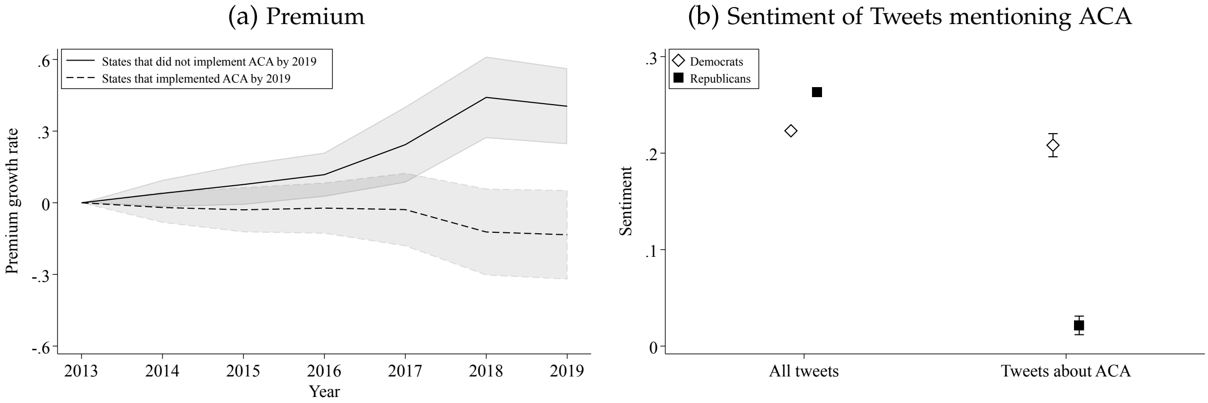


so that the high policy effectiveness grows, one can tell more hopeful stories about them, namely declare a higher $\hat{\mu}_h^H$. More interestingly, the same is true also when the low policy is more effective (μ_l^* is higher) since narratives attribute to the policy they support merits that are actually due to the opposing policy, in practice "stealing effectiveness" from it. Importantly, the simple intuition developed here generalizes to Proposition 2. As anticipated, the comparative statics drives the consequences of our model of narratives for political competition, which we develop in Sections 3 and 4. We conclude the Section with some suggestive evidence supporting it, referring to Appendix D for further details.

Suggestive Evidence We analyze all the tweets shared by U.S. congress members elected between 2012 and 2019, focusing on the debate about the Affordable Care Act (ACA) which has been a flagship policy for the Democrats, proposed as a measure to contain household healthcare expenditure. Restricting attention to this policy is particularly appropriate for our purposes. First, the implementation of the ACA has been staggered (as detailed in Figure A1) across U.S. states and associated with desirable, but delayed variation in insurance premia. Indeed, while premia increased for almost all U.S. states between 2013 and 2019, they increased less in those states which implemented the ACA. We show this in Figure 2a, where we consider an aggregate measure of the premium and we plot its yearly growth rate with respect to 2013, detrended by the average growth across U.S. states[13]. This allows us to credibly state that we are considering a policy which in objective terms is correlated with good outcomes, though this may be difficult to appreciate. Second, the debate on the policy has

[13] The aggregate measure is the average monthly premium calculated using premium and enrollment data for all individual market plans, with data from the Centers for Medicare and Medicaid Services, "Medical Loss Ratio Data and System Resources". In Figure A2 we show that the delay is not driven by late implementation.

Figure 2: Affordable Care Act: premium and sentiment.



(a) Premium

(b) Sentiment of Tweets mentioning ACA

Notes: Panel (a) present the growth rate of an aggregate measure of premium detrended by its national growth rate for states that did not and did implement the ACA by 2019. Shaded areas denote 95 percent confidence intervals. Panel (b) present the average sentiment of all tweets and tweets about ACA separately for Democrats and Republicans. Bars denote 95 percent confidence intervals.

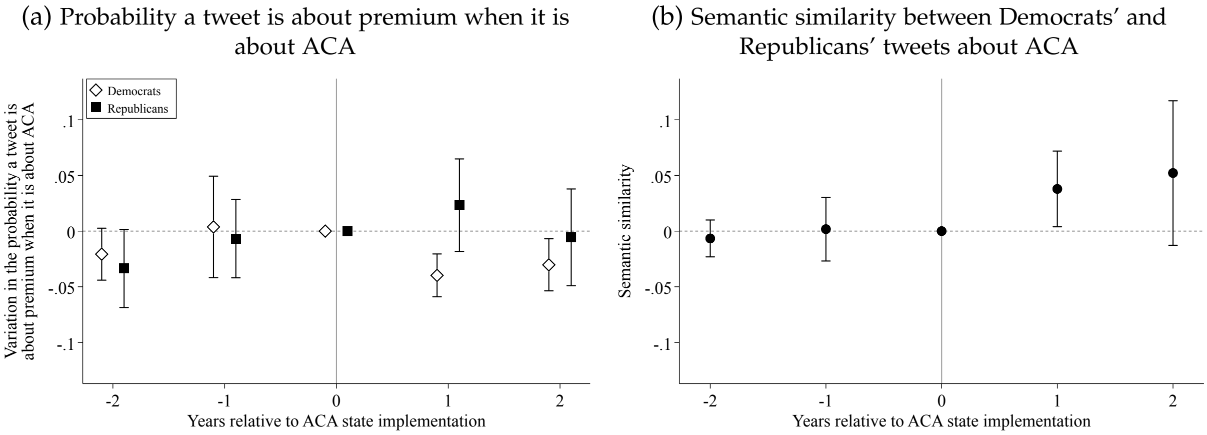
been heated. To confirm this, we performed sentiment analysis on Democrats and Republican tweets, finding, as expected, a marked difference between the groups on average, as shown in Figure 2b. This allows us to credibly state that when Democrats and Republicans share tweets about the topic, they respectively speak positively and negatively about it.

In this context, we find suggestive evidence that (i) Democrats cannot claim credit for their success, while Republicans keep blaming this intervention; (ii) Democrats and Republicans are forced to reduce their disagreement throughout the staggered implementation of the policy. Concerning the first point, through an event study at the politician level exploiting the staggered implementation of the ACA, we observe a significant and sizeable reduction in the probability that Democrats talk about "premium" in association with "ACA", at state level, while this does not happen for Republicans. This is shown in Figure 3a. What is more, we find that this pattern is driven by states where the ACA was *more* effective. Hence, we observe that Democrats tend to stop riding the narrative about the effectiveness of the ACA, in line with the idea that as a policy gets implemented, it becomes less valuable in narrative competition, despite its performance. Concerning the second point, Democrats' and Republicans's tweets about the ACA become closer in a semantic similarity metric throughout the implementation of the Affordable Care Act, as shown in Figure 3b. This points to a reduction in disagreement about the effectiveness of a policy after it gets implemented, as this shrinks the set of narratives deemed plausible.

2 Optimal Narrative Design

In this section, we generalize the problem of designing an optimal narrative to the case of any finite set of policies $\mathcal{A} = \{a_1, \ldots, a_n\}$, any measurable space $\mathcal{Y}$ of outcomes and any

Figure 3: Suggestive evidence related to the Affordable Care Act.



Notes: Panel (a) presents the results of an event study at the politician level describing how the probability a tweet is about premium when it is about ACA varies with the staggered implementation of the Affordable Care Act at the state level within Democrats and within Republicans. Bars denote 95 percent confidence intervals with standard errors clustered at the politician level. Panel (b) presents the results of an event study at the politician level describing how for each Democrat (Republican) the average semantic similarity between his/her tweets and the tweets posted during the same year by all the Republicans (Democrats) elected in their same state varies with the staggered implementation of the Affordable Care Act at the state level. Bars denote 95 percent confidence intervals with standard errors clustered at the state level.

measurable utility $u : \mathcal{Y} \to \mathbb{R}$. Of course, we will mostly have in mind either finite outcome sets or subsets of a Euclidean space. While the main intuitions from the binary case extend, the present analysis reveals some properties which are masked in the binary setting. For instance, it shows that the problem depends only on ordinal rather than cardinal properties of the voter's utility: in particular, risk attitudes play no role. Morever, when $\mathcal{Y} = \mathbb{R}$ and the utility function is increasing, narrative design can be tackled via first order stochastic dominance arguments relating our problem to those in Horowitz and Manski (1995). This case also highlights how the solution to the problem induces a comonotone coupling between the two dimension of the voter's coarse memory: indeed, we can reformulate the problem as one of optimal transport with supermodular matching function (Galichon, 2018, Chap. 4) as shown in Appendix C.1. Finally, increased generality naturally gives the model more flexibility, which may be useful to study further applications.

General Setting Both the *true model* of the economy and any *narrative* are maps denoted $\mu^*, \mu : \mathcal{A} \to \Delta(\mathcal{Y})$, i.e. Markov kernels collecting, respectively, objective and claimed interventional distributions associated to each policy. A *coarse memory* is a pair of probability distributions, one over policies and one over outcomes, $(\alpha, \nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{Y})$. We maintain that the two components are related by μ^*, i.e.

$$\forall O \subseteq \mathcal{Y} \text{ } meas. \int_{\mathcal{A}} \mu^*(a)(O) d\alpha(a) = \nu(O). \tag{7}$$

Moreover, we assume that α has full support[14]. A narrative μ is deemed *plausible* at memory (α, ν) if ν can be expressed as the average of μ with respect to α. In, other words, the memory-dependent set of plausible narratives is

$$\mathcal{M}(\alpha, \nu) = \left\{ \mu : \mathcal{A} \to \Delta(\mathcal{Y}) \;\middle|\; \forall O \subseteq \mathcal{Y} \textit{ meas.} \int_{\mathcal{A}} \mu(a)(O) d\alpha(a) = \nu(O) \right\}.$$

Two techical remarks are in order. First, given the finiteness of $\mathcal{A}$: (i) all the above integrals are actually sums, we choose the more general notation to make it homogeneous; (ii) we abuse notation and write $\alpha(a)$ for $\alpha(\{a\})$. Second, recall that, given any probability measure $\lambda \in \Delta(\mathcal{Y})$, if a probability measure λ' is absolutely continuous with respect to λ (i.e. for any measurable $O \subseteq \mathcal{Y}$, $\lambda(O) = 0 \Rightarrow \lambda'(O) = 0$; notation: $\lambda' << \lambda$) then, by the Radon-Nikodym theorem, λ' admits a density, $\frac{d\lambda'}{d\lambda}$ with respect to λ, i.e. a measurable function $\frac{d\lambda'}{d\lambda} : \mathcal{Y} \to \mathbb{R}_+$ such that for any measurable $O \subseteq \mathcal{Y}$, $\lambda(O) = \int_O \frac{d\lambda'}{d\lambda} d\lambda$.

The Narrative Design Problem To state the generalization of $\mathrm{P}_{\text{simple}}$, we endow the voter with any measurable Bernoulli utility $u : \mathcal{Y} \to \mathbb{R}$, maintaining that she is an anticipatory utility maximizer. The politician's objective is still to promise the most desirable scenario, in anticipatory utility terms, that a plausible narrative can attribute to his policy. Hence, when the voter recalls the distribution of actions $\alpha \in \Delta(\mathcal{A})$ and that of outcomes $\nu \in \Delta(\mathcal{Y})$ the problem of a politician committed to action a is

$$\begin{aligned} V_a(\alpha, \nu) = \max_{\mu : \mathcal{A} \to \Delta(\mathbb{R})} \quad & \mathbb{E}_{\mu(a)}[u] \\ \text{subj. to:} \quad & \mu \in \mathcal{M}(\alpha, \nu) \end{aligned} \tag{$\mathrm{P}_{\text{general}}$}$$

where the notation emphasizes how, according to the narrative μ, the random variable associated with the outcome follows distribution $\mu(a)$ when intervention a is performed.

Our main theorem is a characterization of the solution of $\mathrm{P}_{\text{general}}$ and of its value. Essentially, the result states that the optimal narrative attributes to policy a, by conditioning, all outcomes where u attains a value above a certain threshold, namely a superset of u. Plausibility imposes a bound on the height of such a superset, asking that the outcomes left outside of it have mass at most $1 - \alpha(a)$ according to ν. Hence, the value of the problem is just the expectation of u, conditional on it taking only its "top-$\alpha(a)$-values". Note that every problem for which such a superset is the same will admit the same optimal narrative. In particular, then, any monotone transformation of u implies the same optimal narrative, so that its concavity does not matter for the solution.

[14] This is without essential loss of generality: if one is really interested in the case $\alpha(a) = 0$ one can consider a compactification $\bar{\mathcal{Y}}$ of $\mathcal{Y}$ and restrict to continuous u. Then the problem is solved by the proposal $\hat{\mu}(a) = \delta_{\arg\max u}$

Theorem 1 *Fix a coarse memory $(\alpha, \nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{Y})$. The problem of the politician is solved by any narrative $\hat{\mu} \in \mathcal{M}(\alpha, \nu)$ such that $\hat{\mu}(a)$ has the following density with respect to ν*

$$\frac{d\hat{\mu}(a)}{d\nu} = \frac{1}{\alpha(a)}[\mathbb{1}_{u(y)>\hat{u}} + c\mathbb{1}_{u(y)=\hat{u}}],$$

where $\hat{u} = \inf\{r \mid \nu(\{y \mid u(y) > r\}) \leq \alpha(a)\}$ and c solves $c\nu(\{y|u(y) = \hat{u}\}) = \alpha(a) - \nu(\{y|u(y) > \hat{u}\})$. The value of the problem is

$$V_a(\alpha, \nu) = \frac{1}{\alpha(a)} \int_{\{y|\ u(y)\geq\hat{u}\}} u(y) d\nu(y) = \mathbb{E}_\nu[u|u \geq \hat{u}].$$

Example 1 *To see the connection with the binary case, let us inspect how the Theorem nests the solution to P_{simple} (Equation 4). Recall that $\mathcal{Y} = \{g, b\}$ and $u(g) = 1, u(b) = 0$. In this case, u has two supersets: $\{g, b\}$ (associated to $\hat{u} = 0$) and $\{g\}$ (associated to $\hat{u} = 1$). Consider memories such that $\nu_g > \alpha_h$: clearly, the superset $\{g\}$ is selected, since $\nu(u(y) > r) = \nu(\emptyset) = 0 \leq \alpha_h$ for any $r \geq 1$, but as soon as $r < 1$, $\nu(u(y) > r) = \nu(\{g\}) = \nu_g > \alpha_h$. Hence, in this case $\nu(\{y|u(y) = \hat{u}\}) = \nu_g$ while $\nu(\{y|u(y) > \hat{u}\}) = 0$, and in turn $c = \frac{\alpha_h}{\nu_g}$. This means that $\frac{d\hat{\mu}(s)}{d\nu} = \frac{1}{\nu_g}\mathbb{1}_{u(y)=1} = \frac{1}{\nu_g}\mathbb{1}_{y=g}$ which means $\hat{\mu}(s)(\{g\}) = 1$. For memories such that $\nu_g \leq \alpha_h$ a similar computation implies $\hat{u} = 0$ implying $c = \frac{\alpha_h - \nu_g}{1-\nu_g}$. This means that $\frac{d\hat{\mu}(s)}{d\nu} = \frac{1}{\alpha_h}\left[\mathbb{1}_{y=g} + \frac{\alpha_h-\nu_g}{1-\nu_g}\mathbb{1}_{y=b}\right]$. In turn, this means $\hat{\mu}(s)(\{g\}) = \frac{\nu_g}{\alpha_h}$.*

What can we say about the structure of the optimal narrative on $a' \neq a$? It is easy to note that the politician is indifferent with respect to how the narrative redistributes the worst $1 - \alpha(a)$ outcomes among interventions different from a, and hence many narratives are optimal if $|\mathcal{A}| > 2$. Nonetheless, the distribution of outcome which the narrative associates to the event in which a is *not* implemented is uniquely pin down. For any narrative μ define the distribution $\mu(\neg a) = \frac{1}{1-\alpha(a)} \sum_{a' \neq a} \alpha(a')\mu(a')$, which is the average of the interventional distributions the narrative associates to policies different from a.

Corollary 1 *In any optimal narrative $\hat{\mu}$, $\hat{\mu}(\neg a)$ has density*

$$\frac{d\hat{\mu}(\neg a)}{d\nu} = \frac{1}{1 - \alpha(a)}[\mathbb{1}_{u(y)<\hat{u}} - c\mathbb{1}_{u(y)=\hat{u}}].$$

Hence, any optimal narrative $\hat{\mu}$ induces the same $(\hat{\mu}(a), \hat{\mu}(\neg a))$, which we call a sufficient representation.

All results are proved in Appendices A and B. The Theorem relies on the *bathtub principle* from measure theory (Lieb and Loss, 2001, Theorem 1.14, for instance), while the Corollary is obtained immediately, deriving by ν (in the Radon-Nikodym sense) the plausibility constraint once it is rewritten as

$$\alpha(a)\mu(a) + (1 - \alpha(a))\mu(\neg a) = \nu. \tag{8}$$

The logic of the proof can be understood via first order stochastic dominance arguments in the case where u is increasing. In this subcase one can think of $\mathrm{P}_{\text{general}}$ as the problem of an econometrician who has to find the upper identification bound for the expected utlity of outcome – an increasing functional, cfr. Horowitz and Manski (1995) – conditional on the policy he committed to, given the marginal data (α, ν).

Monotone Increasing Model Consider any strictly increasing $u : \mathcal{Y} \subseteq \mathbb{R} \to \mathbb{R}$ and, let F_ν be the CDF of ν. In this case, by an equivalent characterization of first order stochastic dominance[15], to solve the problem is sufficient to find a plausible narrative μ such that $\mu(a)$ first order stochastic dominates $\mu'(a)$ for any other plausible μ'. From Equation 8, we can intuitively see how plausibility imposes a constant relative cost of $\frac{\alpha(a)}{1-\alpha(a)}$ to improve μ in the FOSD order. Such a cost is constant across y levels and depends only on $\alpha(a)$. Hence, the solution will involve concentrating $\mu(a)$ on higher and higher values, as long as one finds some $\mu(\neg a)$ to complete $\mu(a)$ to a plausible narrative. Indeed, consider starting from the true model μ^*. Then, imagine to modify $\mu^*(a)$ by forcing the corresponding variable to be 0 up to a certain threshold $y' > \inf \operatorname{supp}(\nu)$, obtaining a new distribution $\mu'(a)$. For a sufficiently small y' there will be some $\mu'(\neg a)$ which makes $\mu'(a)$ plausible, which can be computed inverting Equation 8. Note that $\mu'(a)$ first order stochastic dominates $\mu^*(a)$. Imagine repeating over again this process of adjustments: at some point, it will be impossible to satisfy plausibility, reaching the solution to the problem. Mathematically, this corresponds to keeping the CDF $F_{\mu(a)}$ equal to 0 as long as possible. To do this, while respecting plausibility, one has to make the CDFs $F_{\mu(a')}$ for all other interventions $a \neq a'$ grow as fast as possible. Imposing this in Equation 8, rewritten in terms of CDFs, one can show the following restriction of Theorem 1.

Proposition 1 *Fix a coarse memory $(\alpha, \nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{Y})$. In the case where u is increasing, the optimal narrative is determined by the following CDFs:*

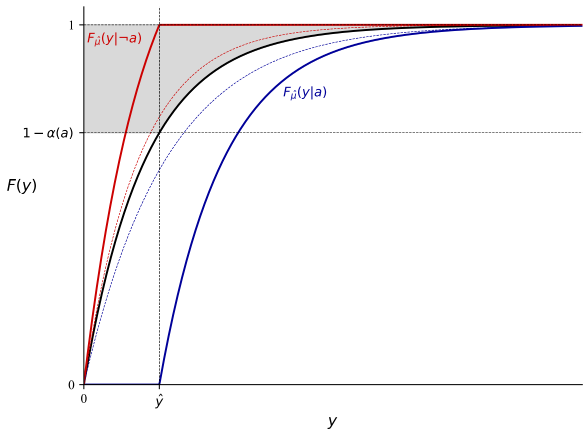
$$\begin{cases} F_{\hat{\mu}(a)} & = \max\left\{\frac{F_\nu - (1-\alpha(a))}{\alpha(a)}, 0\right\} = \frac{F_\nu(y)-(1-\alpha(a))}{\alpha(a)}\mathbf{1}_{y \geq \hat{y}} \\ F_{\hat{\mu}(\neg a)} & = \min\left\{\frac{F_\nu}{1-\alpha(a)}, 1\right\} = \frac{F_\nu(y)}{1-\alpha(a)}\mathbf{1}_{y \leq \hat{y}} + \mathbf{1}_{y > \hat{y}} \end{cases} \tag{9}$$

where $\hat{y} = F_\nu^{-1}(1-\alpha(a))$.

Essentially, the couple $(F_{\hat{\mu}(a)}, F_{\hat{\mu}(\neg a)})$ is determined by a threshold, $\hat{y} = F_\nu(1-\alpha(a))$, such that $F_{\hat{\mu}(a)}$ is the conditional distribution of ν, given that the outcome is above $\hat{y}$, while $F_{\hat{\mu}(a)}$ is the conditional distribution of ν, given that the outcome is below $\hat{y}$. Indeed, $[\hat{y}, \infty)$ is exactly the superset of mass $\alpha(a)$ commanded by Theorem 1. Figure 4 illustrates the construction.

[15] For convenience, we recall that given two distributions X, Y whose supports are included in $\mathcal{Y}$ and whose CDFs are respectively F, G: X *first order stochastically dominates* (FOSD) Y if $F(y) \leq G(y)$ $\forall y$, equivalently, X FOSD Y if and only if, for any increasing u, $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ (Shaked and Shanthikumar, 2007, for a standard reference).

Figure 4: Construction of the Optimal Narrative: Increasing-Continuous



Notes: The black curve is the CDF of ν for an economy where $\alpha(a) = 0.3$ (and $\alpha(\neg a) = 0.7$) and the true model is such that $\mu^*(a) \sim Exp(2)$ (whose CDF is in the dashed blue curve) and $\mu^*(\neg a) \sim Exp(3)$ (whose CDF is in the dashed red curve). The blue solid curve is $F_{\hat{\mu}}(y|a)$, while the red curve is $F_{\hat{\mu}}(y|\neg a)$. Finally, the gray area is the value $\alpha(a) \times V(\alpha, \nu)$ for $u = id$.

The argument sketched here is developed into a self-contained proof for Proposition 1 in Appendix B. We will meet this case again in Section 3, where it will come in handy to illustrate narrative polarization.

Comparative Statics Moving on, we are interested in studying the comparative statics of $\text{P}_{\text{general}}$ with respect to the memory of policies $\alpha \in \Delta(\mathcal{A})$ and to the true model of the economy μ^*, maintaining Equation 7.[16] Accordingly, we will write $\nu \equiv \nu(\alpha, \mu^*)$. Again, our result parallels the analysis of the toy model in Section 1. To this end, we have to define appropriate binary relations on $\Delta(\mathcal{A})$ and on $\Delta(\mathcal{Y})^{\mathcal{A}}$ to replace the standard order on $\mathbb{R}$ adopted in Section 1. For what concerns memories, note that since we are dealing with probability measures, an increase in the frequency of intervention a implies that the frequency of all the others decreases in the *aggregate*, but such an aggregate decrease may be compatible with many different patterns. We are interested in those variations such that, whenever $\alpha(a)$ goes up, $\alpha(a')$ goes down for any $a' \neq a$, since we feel they capture the idea that the frequency of intervention a grows or diminishes *independently* from the frequency of all the other interventions. This will be the case in our application (Section 4) when α evolves as a result of elections, which involve a sole winner. Formally, given $\alpha, \alpha' \in \Delta(\mathcal{A})$ we say that[17]

$$\alpha \textit{ a-majorizes } \alpha' \Leftrightarrow \alpha(a) \geq \alpha'(a) \text{ and } \forall a' \neq a \ \alpha(a') \leq \alpha'(a').$$

[16] Studying the comparative statics with respect to α and ν in isolation would actually much easier, but against our tenet that policy implementation influences, via the true model, that of outcomes (something which is crucial for the dynamic in Section 4).

[17] We choose this name to highlight the link with majorization orders between vectors. Our notion implies that α majorizes α' when coordinates are ordered so that the component relative to policy a is the first.

For what concerns models, we want to formalize the idea that $\mu^{*'}$ is a better economy than μ^* if it produces more good outcome, in the sense defined by u. To capture this, define, for any $r \in \mathbb{R}$ the superset of u of height r, $S(r) = \{y | u(y) \geq r\}$. Then, given $\mu^{*'}, \mu^* \in \Delta(\mathcal{Y})^{\mathcal{A}}$ we say that

$$\mu^{*'} \textit{ is weakly more productive than } \mu^* \Leftrightarrow \forall a \in \mathcal{A},\ r \in \mathbb{R}\ \mu^{*'}(a)(S(r)) \geq \mu^*(a)(S(r)).$$

This says that, fixed a superset, all interventional distributions of $\mu^{*'}$ put at least as mass in it as those of μ^*. Note that this includes the case where $\mu^*(a)$ stays fixed and a *different* policy becomes more productive. Also, note that, whenever u is an increasing function, the definition boils down (weak) first order stochastic dominance of $\mu^{*'}$ on μ^*, policy-wise. Then we are ready to state our result.

Proposition 2 *Fix any $a \in \mathcal{A}$. The following comparative statics holds:*

1. *Fix μ^*. Then, if α' a-majorizes α, $V(\alpha', \nu(\alpha, \mu^*)) \leq V(\alpha, \nu(\alpha, \mu^*))$, i.e. the value of the problem decreases in the majorization relation.*
2. *Fix α. Then, if $\mu^{*'}$ is weakly more productive than μ^*, $V(\alpha, \nu(\alpha, \mu^{*'})) \geq V(\alpha, \nu(\alpha, \mu^*))$, i.e. the value of the problem increases in the productivity relation.*

3 A Static Game of Narrative Competition

In this Section, we study a game where politicians compete on narratives, as anticipated previously. The payoff structure captures the idea that the voter derives anticipatory utility from narratives, conditional on their plausibility. Hence, the voter evaluates the effects of a commitment policy according to the narrative proposed to support it, and elects who advances the most promising one[18]. The assumption that different policies are evaluated according to different narratives is potentially concerning. In Appendix C.2, we show that the predictions of our model do not change if the timing of the game is modified so that, in a first stage, the agent chooses the most promising narrative and in a second she uses both its dimensions to evaluate the two candidates. While we model competition in a stylized fashion, our game provides a first application of our main results: it allows to understand which narratives prevail in equilibrium, setting the stage for the dynamics studied in Section 4, and to understand narrative polarization.

Static Game Retrieving the story begun in Section 1, we consider two politicians H and L, committed respectively to policies h (high) and l (low), whose effect on outcome is de-

[18] This choice criterion between competing narratives is analogous to Eliaz and Spiegler (2020). It can be seen as an extension of the ideas of Bénabou and Tirole (2016) from beliefs to models.

scribed by the true model of the economy $\mu^* : \{h, l\} \to \Delta(\mathcal{Y})$. H and L announce narratives $\mu^H, \mu^L : \{h, l\} \to \Delta(\mathcal{Y})$ and each gets a positive payoff if a voter V cast her preference in his favor. The voter has coarse memory, and recalls $\alpha \in \Delta(\{h, l\})$, which we identify with $\alpha_h \in (0, 1)$, together with $\nu = \alpha_h \mu^*(h) + (1 - \alpha_h)\mu^*(l) \in \Delta(\mathcal{Y})$. Moreover, she evaluates outcomes according to a Bernoulli utility function $u(y)$. After politicians announce their narratives, the voter tests for their plausibility and votes for the proposer of the most promising plausible narrative, up to a uniform popularity shock $\phi \sim \mathcal{U}\left(\left[-\frac{1}{2\zeta}, \frac{1}{2\zeta}\right]\right)$. More precisely, if only one proposed narrative is plausible, the vote is cast in favor of the proposer of such narrative; if no narrative is plausible, the voter votes by tossing a fair coin; when both are plausible, the voter votes for H if

$$\mathbb{E}_{\mu^H(h)}[u(Y)] \geq \mathbb{E}_{\mu^L(l)}[u(Y)] + \phi \qquad \phi \sim \mathcal{U}\left(\left[-\frac{1}{2\zeta}, \frac{1}{2\zeta}\right]\right) \tag{10}$$

and for L otherwise. Note that incentives are such that each candidate's optimal narrative is independent of the opponent's one. Instead, both candidates try to shape the most promising pitch about their policy given the voter's memory. If follows immediately that the game admits a unique Nash equilibrium where each politician announces the solution to $\mathrm{P}_{\text{general}}$. A direct application of Theorem 1 proves the following fact.

Proposition 3 *Fix any coarse memory $(\alpha, \nu) \in \Delta(\{h, l\}) \times \Delta(\mathcal{Y})$. The game admits a unique equilibrium $(\hat{\mu}^H, \hat{\mu}^L)$ where $\hat{\mu}^H$ solves $V_h(\alpha, \nu)$ and $\hat{\mu}^L$ solves $V_l(\alpha, \nu)$.*

Uniqueness follows from the fact that when $\mathcal{A}$ has two elements, sufficient representations of narratives coincide with full ones. Note that in equilibrium, due to the popularity shock, the outcome of the game is stochastic. Clearly, the probability of winning for any candidate is determined by the difference between the anticipatory utility the two candidates can induce in the voter through their narratives. Focusing, without loss, on H we can define the *narrative advantage*

$$\delta(\alpha, \nu) = V_h(\alpha, \nu) - V_l(\alpha, \nu) = \mathbb{E}_\nu[u | u \geq \hat{u}^H] - \mathbb{E}_\nu[u | u \geq \hat{u}^L], \tag{11}$$

where $\hat{u}^H$ and $\hat{u}^L$ identify the supersets on which the two narratives concentrate the outcome of the policy they support, as described in Theorem 1. Hence, the probability that H wins when the memory of the voter is (α, ν) is just a transformation of this quantity via the CDF of the random shock, $F_\phi(y) = \min\left\{\max\left\{\zeta\left(y + \frac{1}{2\zeta}\right), 0\right\} 1\right\}$.

$$P^H(\alpha, \nu) = F_\phi(\delta(\alpha, \nu)).$$

To determine which memories are favorable to H and which are favorable to L, we undertake a simple qualitative study[19] of P^H. To this end, recall that ν is related to α via the true model μ^* as in Equation 7. This allows to regard the distribution of outcome, the narrative advantage and ultimately the probability of winning as functions of (α, μ^*). Since μ^* is fixed throughout, these are ultimately just functions of α_h. Then, we can note that, by point 1 of Proposition 2, as α_h grows, $\delta(\alpha_h)$ diminishes. Moreover, it becomes zero when $\alpha_h = \alpha_l = 0.5$. Combining these observations, we can obtain the following qualitative characterization of the probability of winning.

Proposition 4 *For any μ^*, the following hold:*

1. *The more a policy has been implemented, the less probable its proponent wins: $P^H(\alpha_h)$ is decreasing.*
2. *The mass of memories favorable to each candidate is the same: $P^H(\alpha_h)$ has a fixed point at $\alpha_h = 1/2$.*
3. *When the shock ϕ is sufficiently concentrated[20], there are memories where one of the candidate wins for sure. Formally, there exists $\underline{\delta}, \overline{\delta}$ such that when $\zeta > \max\left\{\frac{1}{2|\underline{\delta}|}, \frac{1}{2|\overline{\delta}|}\right\}$, $P^H(\alpha_h) = 1$ for $\alpha_h \geq \delta^{-1}\left(-\frac{1}{2\zeta}\right)$ and $P^H(\alpha_h) = 0$ for $\alpha_h \leq \delta^{-1}\left(\frac{1}{2\zeta}\right)$.*

Point 2 captures the consequences of the merit-stealing, buck-passing mechanism for the competition game: independently of the true model of the economy, and hence from the true quality of each policy, memories are split equally between those in favor and those against each candidate. While this is an observation about the static model, the existence of this fixpoint implies that in our dynamics (Section 4) the system stays trapped in a state where most narratives are plausible, which is self confirming.

Polarization In the Monotone Increasing Model, the explicit expression for the narratives' CDFs (Proposition 1) allows us to quantify disagreement in terms of a common distance between probability measures, namely the Kolmogorov-Smirnov metric, making precise in which sense equilibrium narratives are polarized in our model. Recall that, given $\lambda, \lambda' \in \Delta(\mathcal{Y}) \subseteq \Delta(\mathbb{R})$, the Kolmogorov-Smirnov metric is defined by

$$d^{KS}(\lambda, \lambda') = \sup_{y \in \mathcal{Y}} |F_\lambda(y) - F_{\lambda'}(y)|,$$

and it captures the largest discrepancy between the two CDFs of the measures, providing an index of how different the distributions are across their entire range of values. From it, we

[19] The qualitative properties of P^H are quite robust and do not rely on the shock being uniform. The first holds for any shock, the second relies on symmetry (around 0), and the last on compactness of the shock's support.

[20] This is always the case, for instance, when $\mathcal{Y} = \mathbb{R}$ and the interventional distributions of the true model μ^* have unbounded support.

can define the following notion of *distance between narratives*[21]

$$d^{\mathcal{M}}(\mu, \mu') = \frac{1}{2}[d^{KS}(\mu(h), \mu'(h)) + d^{KS}(\mu(l), \mu'(l))].$$

Then we have the following result, which generalizes the intuition about polarization developed in Section 1.

Proposition 5 *Consider the monotone increasing model. Fix any $\alpha \in \Delta(\mathcal{A})$. The equilibrium narratives $(\hat{\mu}^H, \hat{\mu}^L)$ maximise $d^{\mathcal{M}}(\mu, \mu')$ over $\mathcal{M}(\alpha, \nu(\alpha, \mu^*))$. Moreover, for any μ^* with continuous interventional distributions, $d^{\mathcal{M}}(\hat{\mu}^H, \hat{\mu}^L)(\alpha_h)$ is maximised at $\alpha_h = \frac{1}{2}$.*

The result follows from the structure of optimal narratives in the monotone increasing case (Proposition 1) which ensures that politician A associates to policy a an interventional distribution which FOSD any other plausible interventional distribution for that policy. Hence, the first point follows since, for any memory $\alpha \in \Delta(\{h, l\})$, optimal narratives move outcome mass in opposite directions. The second point follows since the more some policy a has been implemented in the voter's memory, the more plausible narratives must be realistic on that dimension. Since this holds for *both* politicians, they are forced to converge about the effectiveness of policy a. Balanced memories are those where such convergence is minimized. We now turn to the study of a dynamic extension of our game, which reveals how these are precisely the memories to which the system converges, asymptotically.

4 Dynamics

In this section, we consider repetitions of the static game described above, for discrete times $\tau = 1, 2, \dots$. Following a simplification that is common to many political economy models, we assume that politicians are myopic, so that at each period they play narratives which are an equilibrium of the static game described in Section 3. The stochastic winner implements his commitment policy, affecting the distribution of outcome, as dictated by the true model of the economy. The coarse memory of the voter is updated accordingly, in particular, we assume that memory tracks the *historical frequency* of implemented policies and realized outcomes. This is a natural choice, which allows to isolate the effects of coarseness on the dynamics, abstracting from other biases, such as limitation in length. As intuitive, cycles of political power arise in the model. Perhaps more surprisingly, we discover that both politicians hold office with the same frequency in the long run, independently of the objective quality of their commitments. This is a consequence of the fixpoint property for the probability of winning in the static model (Point 2 in Proposition 4) and a clear consequence of the

[21] Regarding narratives as pairs of probability measures, up to the $\frac{1}{2}$ factor, we are just considering the 1-product metric on the product of two copies of the space of probability distributions endowed with the Kolmogorov-Smirnov metric. Indeed, $d^{\mathcal{M}}(\mu, \mu') = \frac{1}{2}||(d^{KS}(\mu(h), \mu'(h)), d^{KS}(\mu(l), \mu'(l)))||_1$

free riding effect induced by narratives in the dynamic context. The analysis is complemented by Appendix C.3 where we consider alternative laws of motions showing how, if ϕ is concentrated enough, political cycles are a robust prediction of our model. Therein, we also inspect the dynamics in the case where memory, beyond coarseness, is limited in length, so that it follows a *moving average* with persistence $\kappa \in (0,1)$, capturing *recency bias*: again, asymptotic properties are essentially the same.

Timing The timing is as follows. Given an initial memory[22] $\alpha_h^0 \in (0,1)$, at each time τ:

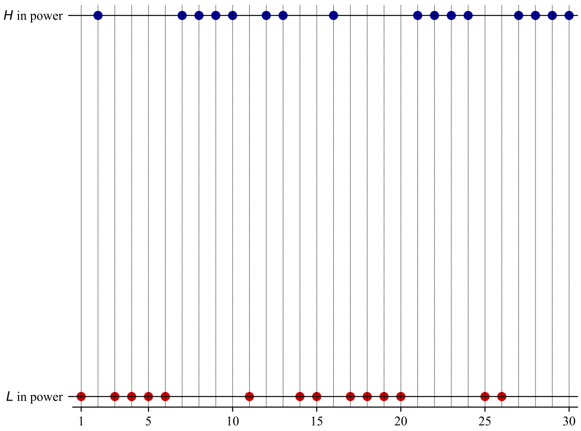
1. Politicians announce narratives $\hat{\mu}^{\tau,H}(\alpha^\tau, \mu^*)$ and $\hat{\mu}^{\tau,L}(\alpha^\tau, \mu^*)$.
2. A random popularity shock $\phi^\tau \sim \mathcal{U}\left[-\frac{1}{2\zeta}, \frac{1}{2\zeta}\right]$, where $\zeta > 0$, affects L's popularity, independently of previous shocks realizations.
3. The representative voter casts her vote, determining time τ winner.
4. The winner implements his commitment policy (h for H and l for L) which we code as a boolean variable $w^\tau \in \{0 \equiv l, 1 \equiv h\}$.
5. The voter's memory evolves according to

$$\begin{cases} \alpha_h^{\tau+1} & = \frac{\tau+1}{\tau+2}\alpha_h^\tau + \frac{1}{\tau+2}w^\tau \\ \nu^{\tau+1} & = \alpha_h^{\tau+1}\mu^*(h) + (1-\alpha_h^{\tau+1})\mu^*(l). \end{cases}$$

To understand political competition in the long run, it is sufficient to understand the asymptotic behavior of the stochastic process $(\alpha_h^\tau)_{\tau=0}^\infty$ tracking the frequency of policies' implementation. To develop intuition, consider the binary example of Section 1. Start from some period τ where H is in power and assume the impact of the shock is negligible. Implementation of h raises its frequency in the voter memory so that $\alpha_h^{\tau+1} > \alpha_h^\tau$. In turn, this implies that plausible narratives change, allowing to attribute more good outcomes to policy t and less to policy h. Then, the incumbent is forced to revise his narrative about his own policy to some effectiveness $\mu_h^{H,\tau+1} \leq \mu_h^{H,\tau}$. If this level is high enough, the incumbent stays in power and the same happens for another period. Then, as times goes by, α_h keeps increasing and μ_h^H tends to μ_h^* while μ_l^L tends to 1, so that there must be some $\tau' > \tau$ when $\mu_h^{H,\tau'} < \mu_l^{L,\tau'}$ and a policy shift takes place. Then the process repeats for L. This suggests that α_h will not converge to 0 nor to 1 and political cycles will arise. But more should hold: since the steps of the process decrease in length as $1/t$, we expect that the process converges somewhere. The most natural

[22] Starting from an internal frequency is essentially without loss. Indeed, suppose the first winner is determined by tossing a fair coin. Let's say it's L. Then if ϕ is sufficiently concentrated, H wins with probability one in the first period and α_h^1 is internal. If ϕ is not sufficiently concentrated, H has still a positive probability of winning, and hence, eventually wins. As a consequence, almost surely, there is a finite T such that α_h^T is internal.

Figure 5: Simulation



Notes: Last 30 periods out of 2000 for a simulation with $u = id$ and an exponential parametrization with $\mu^*(h) = \exp(2)$ and $\mu^*(l) = \exp(3)$.

candidate for such a limit point is $\alpha_h^\infty = 1/2$. Indeed, given the shape of P^H, perturbations to this values tend to be rebalanced.

Theorem 2 *For any μ^*, it holds that*

$$\alpha_h^\tau \xrightarrow{p} \frac{1}{2}.$$

In particular, both H and L rise to power infinitely often, and asymptotically govern with the same frequency.

The proof is based on identifying an appropriate subprocess of $(\alpha_h^\tau)_{\tau=0}^\infty$ and then exploiting Doob's Optional Stopping Theorem to bound the probability it departs from $\frac{1}{2}$ asymptotically. A simulation of the process can be visualised in Figure 5. We can observe that, as a consequence of the fact that the asymptotic frequency of government is the same for both politicians, but memory updates are slower and slower, cycles tend to have increasing length. Moreover, away from the steady state, changes in power tend to happen when the incumbent has ridden his favored policy long enough that his narratives approximate the truth about its effectiveness. Finally, at the steady state, the economy is trapped in a situation where polarization is maximal and plausibility has the least grip.

5 Conclusion

In this paper, we proposed a novel definition of narratives as rhetoric devices that link policies and outcomes, stipulating possibly fallacious attributions of the latter to the former. We formulated a notion of plausibility for narratives, which amounts to non-distortion of the marginal information about these two aspects of the economy or, put differently, to observational equivalence to the true model. We argued that plausible narratives may be produced and shared in an economy where voters have difficulties in perceiving correct (causal) links between variables, while being correctly informed about their distributions. Then, we formalized the problem of strategically designing a narrative aimed at defending a commitment policy, and corroborated its comparative statics with an empirical analysis of tweets from U.S. congress members. Finally, we put the framework at work in a model where a boundedly rational voter assesses the plausibility of claims retrospectively, and selects those that she finds more optimistic. We showed how, in such a model, politicians' narratives maximally disagree about the effectiveness of each policy, within the bounds imposed by plausibility. Nonetheless, they are forced to converge on the effectiveness of a policy when it has been implemented sufficiently often. In a dynamic setting, we showed how this mechanism has the potential to generate political cycles.

We foresee further exploration of our framework in future work. On the theory side, our model of narratives may shed light on entry behavior capturing why new parties may be attracted to commit to inferior, but rarely attempted policies on which an effective narrative may be spinned. Moreover, since narratives expropriate politicians from the good outcome they produce, they may drive the most capable types away from politics, adding to the motives of Caselli and Morelli (2004). Relatedly, whereas in the present work the comparison between interventions is based on the upper identification bound for their effectiveness, we may adopt a choice criterion a la Hurwicz (1951) to account for the lower identification bound, too. On the empirical side, we recognize that the present work is just a first stab and further analysis should be carried out. A natural experimental question is to investigate our coarseness assumption and to distinguish it from correlation neglect. In turn, a design similar to that of Barron and Fries (2023) may be adopted to explore the link between coarseness and plausibility experimentally. The same link may be gauged in survey data, adapting methods from Andre et al. (2022), or in text-data, exploiting state of the art techniques (Ash et al., 2023, for instance).

Bibliography

Aina, C. (2021). Tailored stories. Technical report, Mimeo.

Ambuehl, S. and H. C. Thysen (2024). Choosing between causal interpretations: An experimental study. *NHH Dept. of Economics Discussion Paper* (07).

Andre, P., C. Pizzinelli, C. Roth, and J. Wohlfart (2022). Subjective models of the macroeconomy: Evidence from experts and representative samples. *The Review of Economic Studies 89*(6), 2958–2991.

Ash, E., S. W. Mukand, and D. Rodrik (2023). Economic interests, worldviews, and identities.

Barron, K. and T. Fries (2023). Narrative persuasion. Technical Report SP II 2023-301, WZB Discussion Paper.

Barron, K. and T. Fries (2024). Narrative persuasion: A brief introduction encyclopedia of experimental social science.

Battigalli, P. and N. Generoso (2021). Information flows and memory in games. *Available at SSRN 4435785*.

Bayer, C., K. Wälde, et al. (2011). Existence, uniqueness and stability of invariant distributions in continuous-time stochastic models. *Gutenberg School of Management and Economics: Discussion Paper Series*.

Bellodi, L., M. Morelli, A. Nicolo, and P. Roberti (2023). The shift to commitment politics and populism: Theory and evidence. *BAFFI CAREFIN Centre Research Paper* (204).

Bénabou, R. and J. Tirole (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives 30*(3), 141–164.

Bordalo, P., N. Gennaioli, and A. Shleifer (2020). Memory, attention, and choice. *The Quarterly Journal of Economics 135*(3), 1399–1442.

Caselli, F. and M. Morelli (2004). Bad politicians. *Journal of public economics 88*(3-4), 759–782.

DellaVigna, S. and W. Kim (2022). Policy diffusion and polarization across us states. Technical report, National Bureau of Economic Research.

Eliaz, K. and R. Spiegler (2020). A model of competing narratives. *American Economic Review 110*(12), 3786–3816.

Eliaz, K., R. Spiegler, and S. Galperti (2023, January). False Narratives and Political Mobilization. CEPR Discussion Papers 17832, C.E.P.R. Discussion Papers.

Esponda, I. and D. Pouzo (2017). Conditional retrospective voting in large elections. *American Economic Journal: Microeconomics 9*(2), 54–75.

Eyster, E. and G. Weizsacker (2016). Correlation neglect in portfolio choice: Lab evidence. *Available at SSRN 2914526*.

Fréchette, G., E. Vespa, and S. Yuskel (2024). Extracting models from data sets: An experiment. *Working Paper*.

Fudenberg, D., G. Lanzani, and P. Strack (2022). Selective memory equilibrium. *Available at SSRN 4015313*.

Galichon, A. (2018). *Optimal transport methods in economics*. Princeton University Press.

Hairer, M. (2021). Notes on convergence.

Hayek, F. (1974). The pretense of knowledge. Nobel Prize Lecture.

Horowitz, J. L. and C. F. Manski (1995). Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, 281–302.

Horz, C. and K. Kocak (2022). How to keep citizens disengaged: Propaganda and causal misperceptions.

Hurwicz, L. (1951). Some specification problems and applications to econometric models. *Econometrica 19*(3), 343–344.

Hutto, C. and E. Gilbert (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Volume 8, pp. 216–225.

Ispano, A. (2023). The perils of a coherent narrative. Technical report.

Izzo, F., G. J. Martin, and S. Callander (2021). Ideological competition. *American Journal of Political Science*.

Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Levy, G., R. Razin, and A. Young (2022). Misspecified politics and the recurrence of populism. *American Economic Review 112*(3), 928–962.

Lieb, E. H. and M. Loss (2001). *Analysis*, Volume 14. American Mathematical Soc.

Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press.

Manski, C. F. (2004). Social learning from private experiences: the dynamics of the selection problem. *The Review of Economic Studies 71*(2), 443–458.

Nannestad, P. and M. Paldam (2003). The cost of ruling: A foundation stone for two theories. In *Economic voting*, pp. 17–44. Routledge.

Oaksford, M. and N. Chater (2020). New paradigms in the psychology of reasoning. *Annual review of psychology 71*, 305–330.

Ortoleva, P. and E. Snowberg (2015). Overconfidence in political behavior. *American Economic Review 105*(2), 504–535.

Paldam, M. (1986). The distribution of election results and the two explanations of the cost of ruling. *European Journal of Political Economy 2*(1), 5–24.

Pearl, J. (2009). *Causality*. Cambridge university press.

Reimers, N. and I. Gurevych (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American psychologist 54*(3), 182.

Schwartzstein, J. and A. Sunderam (2021). Using models to persuade. *American Economic Review 111*(1), 276–323.

Shaked, M. and J. G. Shanthikumar (2007). *Stochastic orders*. Springer.

Sheingate, A. D. (2016). *Building a business of politics*. Oxford University Press.

Shiller, R. J. (2017). Narrative economics. *American economic review 107*(4), 967–1004.

Spiegler, R. (2016). Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics 131*(3), 1243–1290.

Spiegler, R. (2020). Behavioral implications of causal misperceptions. *Annual Review of Economics 12*, 81–106.

The Guardian (2022). Republicans and democrats flood us midterms with political ads. https://www.theguardian.com/us-news/2022/nov/03/republicans-democrats-political-ads-us-midterms. Accessed May 11, 2023.

A Proofs of the Main Results

The Appendix contains the Proofs of Theorems 1 and 2. The other results are proved in Appendix B.

Proof of Theorem 1 We break the proof in two Lemmata that allow to apply the bathtub principle.

Lemma A1 *Let $\mu \in \mathcal{M}(\alpha, \nu)$. For any $a \in \mathcal{A}$, $\mu(a) << \nu$.*

Proof. By contrapositive, if there exists a measurable subset $O \subseteq \mathcal{Y}$ and some $a \in \mathcal{A}$ such that $\mu(a)(O) > 0$ but $\nu(O) = 0$, then μ cannot be plausible, since

$$\int_{\mathcal{A}} \mu(a)(O) d\alpha(a) = \sum_{a' \in \mathcal{A}} \alpha(a')\mu(a')(0) \geq \alpha(a)\mu(a)(O) > 0 = \nu(O).$$

□

As a consequence, by the Radon-Nykodim Theorem, then, any plausible narrative's interventional distribution admits a density with respect to ν. The next Lemma bounds this density.

Lemma A2 *For any $\mu \in \mathcal{M}(\alpha, \nu)$, $a \in \mathcal{A}$, it holds $\nu - a.e.$ $\frac{d\mu(a)}{d\nu}(y) \leq \frac{1}{\alpha(a)}$.*

Proof. By contrapositive, if there exists a measurable subset $O \subseteq \mathcal{Y}$ and some $a \in \mathcal{A}$ such that $\forall y \in O$ $\frac{d\mu(a)}{d\nu}(y) > \frac{1}{\alpha(a)}$, then

$$\mu(a)(O) = \int_O \frac{d\mu(a)}{d\nu}(y) d\nu(y) > \frac{\nu(O)}{\alpha(a)}.$$

Rearranging, we see that μ cannot be plausible. Indeed,

$$\sum_{a \in \mathcal{A}} \alpha(a)\mu(a)(O) \geq \alpha(a)\mu(a)(O) > \nu(O).$$

□

The following is the main ingredient of the proof (Lieb and Loss, 2001)

Lemma A3 (Bathtub Principle) *Let (Ω, Σ, γ) be a measure space and let $f : \Omega \to \mathbb{R}$ be a measurable function on Ω such that $\gamma(\{x | f(x) < r\})$ is finite for all $t \in \mathbb{R}$. Let the number $G > 0$ be given and define a class of measurable functions Ω by*

$$C(G, \gamma) = \left\{ g | \ \gamma - a.e. \ x \ 0 \leq g(x) \leq 1, \ \int g(x) d\gamma(x) = G \right\}. \tag{12}$$

Then the minimization problem $I = \inf_{g \in C} \int_{\Omega} f(x)g(x)d\gamma(x)$ is solved by

$$g(x) = \mathbb{1}_{(f<s)}(x) + c\mathbb{1}_{(f=s)}(x)$$

where $s = \sup\{r|\ \mu((x : f(x) < r)) \leq G\}$ and $c\mu(\{x|\ f(x) = s\}) = G - \mu(\{x|\ f(x) < s\})$.

To conclude the proof of the Theorem, consider the set of functions $F = \left\{\alpha(a)\frac{d\mu(a)}{d\nu} \mid \mu \in \mathcal{M}(\alpha, \nu)\right\}$. Consider any $g = \alpha(a)\frac{d\mu(a)}{d\nu} \in F$. Note that:

1. It holds

$$\int_{\mathcal{Y}} g d\nu(y) = \int_{\mathcal{Y}} \alpha(a)\frac{d\mu(a)}{d\nu} d\nu(y) = \alpha(a)\mu(a)(\mathcal{Y}) = \alpha(a),$$

   where the last passage follows from $\mu(a)$ being a probability distribution.

2. Since $0 \leq \frac{d\mu(a)}{d\nu} \leq \frac{1}{\alpha(a)}$

$$0 \leq \alpha(a)\frac{d\mu(a)}{d\nu} \leq 1.$$

Hence, we can identify F with $C(\alpha(a), \nu)$. Now, observe that we can rewrite our objective as

$$\min_{\{\mu(a)|\mu \in \mathcal{M}(\alpha,\nu)\}} \int_{\mathcal{Y}} -u d\mu(a)(y) \equiv \min_{\{\frac{d\mu(a)}{d\nu}|\mu \in \mathcal{M}(\alpha,\nu)\}} \int_{\mathcal{Y}} -u\frac{d\gamma|a}{d\nu} d\nu \equiv \min_{g \in C_{\alpha(a)}} \frac{1}{\alpha(a)} \int -u g d\nu.$$

To conclude, since u is measurable and ν, being a probability measure, is a finite measure, we apply the principle with $f \equiv -u, \quad \gamma \equiv \nu, \quad G = \alpha(a)$ and, to obtain the final statement, we just note that

$$\begin{gathered} -u(y) < \sup\{t|\ \nu(\{t| - u(y) < t\}) \leq \alpha(a)\} \Leftrightarrow \\ u(y) > \inf\{r|\ \nu(\{r|u(y) > r\}) \leq \alpha(a)\}. \end{gathered}$$

The Corollary follows by linearity of the Radon-Nikodym theorem, differentiating the plausibility constraint with respect to ν, once it is rewritten as $\alpha(a)\mu(a) + (1 - \alpha(a))\mu(\neg a) = \nu$. Indeed, one has

$$\begin{aligned} \alpha(a)\frac{d\mu(a)}{d\nu} + (1 - \alpha(a))\frac{d\mu(\neg a)}{d\nu} = \mathbb{1}_{\mathcal{Y}} \Rightarrow \frac{d\mu(\neg a)}{d\nu} &= \frac{1}{1 - \alpha(a)}[\mathbb{1}_{\mathcal{Y}} - (\mathbb{1}_{u(y)>\hat{u}} + c\mathbb{1}_{u(y)=\hat{u}})] = \\ &= \frac{1}{1 - \alpha(a)}[\mathbb{1}_{u(y)<\hat{u}} - c\mathbb{1}_{u(y)=\hat{u}}], \end{aligned}$$

where the implication follows substituting the expression for $\frac{d\mu(a)}{d\nu}$.

Proof of Theorem 2 To prove the theorem we show that

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(\alpha_h^\tau < \frac{1}{2} - \varepsilon\right) \to 0 \quad as \quad \tau \to \infty.$$

To show that $\mathbb{P}(\alpha_h^\tau > \frac{1}{2} + \varepsilon)$ also tends to 0, one can adopt a mirror argument. Fix some $\varepsilon > 0$. To begin with, note that it is sufficient to restrict to the study of the subprocess defined by the indices $\mathcal{I} = \{\tilde{\tau} \mid \alpha_h^{\tilde{\tau}} \leq \frac{1}{2} - \varepsilon/2\}$. Note, preliminarly, that, since P^H is decreasing, restricting to $\mathcal{I}$, the process jumps to the left with probability $P^H(\alpha_h^{\tilde{\tau}}) \geq P^H(1/2 - \varepsilon/2) \geq 1/2$. Given this, the first fact we can establish is the following.

Lemma A4 *$(\alpha_h^\tau)_{\tau \in \mathcal{I}}$ is a submartingale.*

Proof. Observe that $\mathbb{E}[\alpha^{\tau+1} | \alpha^\tau] = (1 - P^H(\alpha^\tau))\frac{\tau+1}{\tau+2}\alpha_\tau + P^H(\alpha^\tau)(\frac{\tau+1}{\tau+2}\alpha_\tau + \frac{1}{\tau+2}) \geq \alpha_\tau$ if and only if $P(\alpha_\tau) \geq \alpha_\tau$. This is always true if $\alpha_h^{\tilde{\tau}} \leq \frac{1}{2}$, as in our case. □

Moreover, we can show that anytime α_h^τ ends up in the region defined by $\mathcal{I}$, it evades from it. Formally:

Lemma A5 *Let $\tilde{\tau} \in \mathcal{I}$, there exists $t > \tilde{\tau}$ such that $t \notin \mathcal{I}$.*

Proof. By contradiction, assume that for all $t \geq \tilde{\tau}$, $t \in \mathcal{I}$. Then, the random variable which describes the position of the process at any subsequent time is $A_T = \frac{\tilde{\tau}\alpha_{\tilde{\tau}} + \sum_{i=0}^{T-1} W_{\tilde{\tau}+i}}{\tilde{\tau}+T}$ where $W_{\tilde{\tau}+i} \sim \text{Bernoulli}(P^H(\alpha_h^{\tilde{\tau}+i}))$ describes time $\tilde{\tau} + i$ winner. Note that, while $(W_{\tilde{\tau}+i})_{i=0}^{T-1}$ are not independent, they are independent conditional on the vector of their parameters, whose componets, given our preliminary observation, are all at least $P^H(1/2 - \varepsilon/2) \geq 1/2$. Hence, defining an auxiliary vector of independent Bernoullis $B_{\tilde{\tau}+i}$ of parameter $P^H(1/2 - \varepsilon/2) \geq 1/2$, we have that, for any T, $\mathbb{E}[A_T] \geq \frac{\tilde{\tau}\alpha_{\tilde{\tau}} + \sum_{i=0}^{T-1} \mathbb{E}[B_{\tilde{\tau}+i}]}{\tilde{\tau}+T}$. Hence, taking limits

$$\lim_{T \to \infty} \mathbb{E}[A_T] \geq P^H(1/2 - \varepsilon/2) \geq 1/2 > \frac{1}{2} - \varepsilon/2.$$

Since A_T has vanishing variance, we can conclude by Chebychev's inequality that, with probability 1, there exists some T' such that $A_{T'} > \frac{1}{2} - \varepsilon/2$, obtaining a contradiction. □

Moving on with the proof, observe that, if $\mathcal{I}$ is bounded above, we are done. If this is not the case, let $\tilde{\tau}_0$ be the smallest index in $\mathcal{I}$ higher than $\frac{2N}{\varepsilon}$. By Lemma A5, there exists $\tau_1 \geq \tilde{\tau}_0$ such that $\alpha_h^{\tilde{\tau}_1} \geq 1/2 - \varepsilon/2$. Skip to the smallest $\tilde{\tau}_2 \geq \tau_1$ such that $\tilde{\tau}_2 \in \mathcal{I}$. Observe that it must be $\alpha_h^{\tilde{\tau}_2} \geq 1/2 - \varepsilon/2 - \frac{1}{\tilde{\tau}_2}$. Now, since the process is a submartingale, we can show the following fact.

Lemma A6 *Let $\tilde{\tau}^{up} = \min\{\tau \geq \tilde{\tau}_2 | \alpha_h^\tau \geq 1/2 - \varepsilon/2\}$ and $\tilde{\tau}^{down} = \min\{\tau \geq \tilde{\tau}_2 | \alpha_h^\tau \leq 1/2 - \varepsilon\}$. Then $p = \mathbb{P}(\tau^{down} < \tau^{up}) \leq \frac{2}{\varepsilon \tilde{\tau}_2}$.*

Proof. Define the stopping time $\tilde{\tau}^{stop} = \min\{\tilde{\tau}^{down}, \tilde{\tau}^{up}\}$. By Doob's Optional Stopping Theorem, we have the central inequality below:

$$p(1/2 - \varepsilon) + (1 - p)(1/2 - \varepsilon/2) = \mathbb{E}[\alpha_h^{\tau,\tilde{s}top}] \geq \alpha_h^{\tilde{\tau}_2} \geq 1/2 - \varepsilon/2 - \frac{1}{\tau_2}.$$

Simplifying one obtains the result. □

To conclude, note that then, for every $\tilde{\tau}_3 \geq \tilde{\tau}_2$,

$$\mathbb{P}(\alpha_h^{\tilde{\tau}_3} \leq 1/2 - \varepsilon) \leq \frac{2}{\varepsilon\tau_2} \leq \frac{2}{\varepsilon\tau_0} \leq \frac{1}{N}.$$

Which we can make as small as wanted by choosing a sufficiently high N.

B For Online Publication – Additional Proofs

Proof of Proposition 1 Preliminarily, observe that the plausibility constraint rewrites in terms of CDFs as

$$\forall y \in \mathcal{Y} \qquad \alpha(a)F_{\mu(a)}(y) + (1-\alpha(a))F_{\mu(\neg a)}(y) = F_{\nu}(y). \tag{13}$$

We divide the proof in four steps.

The first step is to check that $\hat{\mu}$ is a plausible narrative. We start by verifying that $F_{\mu(\hat{a})}(y)$, defined by Equation 9 is a CDF. To see this, observe that F_{ν} is a CDF, hence it is weakly increasing and right continuous. It follows that the first term in the max is weakly increasing and right continuous. The same trivially holds for 0, which is the second term in the max. Since the max of two weakly increasing and the right continuous functions is also weakly increasing and the right continuous, we can conclude that $F_{\hat{\mu}(a)}(y)$ has these two properties. Finally, we note that $\lim_{y\to\underline{y}} F_{\mu(\hat{a})}(y) = \max\{-\frac{1-\alpha(a)}{\alpha(a)}, 0\} = 0$, and $\lim_{y\to\overline{y}} F_{\mu(\hat{a})}(y) = \max\{1, 0\} = 1$. A completely analogous reasoning leads to the conclusion that $F_{\hat{\mu}(\neg a)}(y)$ is a CDF. Hence, we can define a narrative $\hat{\mu}$ according to which $\hat{\mu}(a)$ has CDF $F_{\hat{\mu}}$ and $\hat{\mu}(a')$ has CDF $F_{\hat{\mu}(\neg a)}(y)$ for any $a' \neq a$. To see that $\hat{\mu}$ is plausible, just exploit the writing of $F_{\mu(\hat{a})}(y)$ and $F_{\hat{\mu}(\neg a)}(y)$ in terms of F_{ν} and check Equation 13.

The second step is to check that, among all plausible narratives μ, $\hat{\mu}$ is such that $\hat{\mu}(a)$ first order stochastic dominates $\mu(a)$. To see this, pick any $y \in \mathcal{Y}$ and consider two complementary cases. If $y \leq \hat{y} = F_{\nu}^{-1}(1-\alpha(a))$, trivially, it holds that

$$F_{\mu(\hat{a})}(y) = 0 \leq F_{\mu}(y|a).$$

If instead $y > \hat{y}$, it holds that

$$\begin{aligned} F_{\mu(\hat{a})}(y) = \frac{F_{\nu}(y) - (1-\alpha(a))}{\alpha(a)} &= \frac{(\alpha(a)F_{\mu}(y|a) + (1-\alpha(a))F_{\mu}(y|\neg a)) - (1-\alpha(a))}{\alpha(a)} = \\ &= F_{\mu}(y|a) - \frac{1-\alpha(a)}{\alpha(a)}[1 - F_{\mu}(y|\neg a)] \leq F_{\mu}(y|a). \end{aligned}$$

where the first equality follows from Equation 9, the second by definition of plausibility, as in Equation 13, the third is algebra, and the inequality by noting that the square parenthesis is always positive, being $F_{\mu}(y|\neg a)$ a CDF. Hence, we have shown that

$$\forall \mu \in \mathcal{M}(\alpha, \mu^{*}) \forall y \in \mathcal{Y} \qquad F_{\hat{\mu}(a)}(y) \leq F_{\mu}(y|a).$$

The third step is just to note that, by a well known equivalent charachterization of first order stochastic dominance (see, for instance, Shaked and Shanthikumar (2007)), fixed an arbitraty increasing $u : \mathcal{Y} \to \mathbb{R}$,

$$\hat{\mu}(a) \succeq_{FOSD} \mu(a) \Longrightarrow E_{\hat{\mu}(a)}[u(Y)] \geq \mathbb{E}_{\mu(a)}[u(Y)].$$

This step concludes the proof that $\hat{\mu}$ is optimal.

Proof of Proposition 2 For point (a) we show that, for any $g \in C(\alpha(a), \nu(\alpha, \mu^*)$ (defined in Equation 12) there exists a $g' \in C(\alpha'(a), \nu(\alpha', \mu^*))$ such that $g' \leq g$. In light of the proof of Theorem 1, this is sufficient to conclude[23]. Let $g \in C(\alpha(a), \nu(\alpha, \mu^*)$. and observe that it must be $g \gneqq 0$, since $\alpha(a) > 0$. Determine $c \in \mathbb{R}$ such that $\int_{\mathcal{Y}} cgd\nu(\alpha', \mu^*) = \alpha'(a)$, i.e. set $c = \frac{\alpha'(a)}{\int_{\mathcal{Y}} gd\nu(\alpha', \mu^*)}$. We want to show $c \in [0, 1]$. Clerly, $c > 0$, as it is the ratio of positive quantities. Suppose by contradiction that $c > 1$. Equivalently, using $\int_{\mathcal{A}} \mu^*(a) d\alpha(a) = \nu$,

$$\begin{aligned}
\alpha'(a) > \int_{\mathcal{Y}} gd\nu(\alpha', \mu^*) = \sum_{a' \in \mathcal{A}} \alpha'(a') \int_{\mathcal{Y}} gd\mu^*(a') \Leftrightarrow \\
\alpha'(a) \left[1 - \int_{\mathcal{Y}} d\mu^*(a) \right] > \sum_{a' \neq a} \alpha'(a') \int_{\mathcal{Y}} gd\mu^*(a') \Leftrightarrow \\
\sum_{a' \neq a} \left(\frac{\alpha'(a)}{\alpha(a)} \alpha(a') - \alpha'(a') \right) \int_{\mathcal{Y}} gd\mu^*(a') > 0,
\end{aligned}$$

where in the third line we use the fact that, since $g \in C(\alpha(a), \nu(\alpha, \mu^*))$, using $\int_{\mathcal{A}} \mu^*(a) d\alpha(a) = \nu$,

$$\int_{\mathcal{Y}} gd\nu(\alpha, \mu^*) = \alpha(a) \Rightarrow \left[1 - \int_{\mathcal{Y}} gd\mu^*(a) \right] = \frac{1}{\alpha(a)} \sum_{a' \neq a} \alpha(a') \int_{\mathcal{Y}} gd\mu^*(a').$$

To reach a contradiction consider the last line in the above chain of equivalences. Observe that, by definition of a-majorization (i) $\frac{\alpha'(a)}{\alpha(a)} \leq 1$ and hence (ii) $\alpha'(a') \geq \alpha(a') \geq \frac{\alpha'(a)}{\alpha(a)} \alpha(a')$, so that the parenthesis is ≤ 0 while the integrals $\int_{\mathcal{Y}} gd\mu^*(a')$ are all bounded in $[0, 1]$, since g is and $\mu^*(a')$ is a probability measure. Then, we can conclude, considering $g' = cg$. By construction of c, it holds that $\int_{\mathcal{Y}} g' d\nu(\alpha', \mu^*) = \alpha'(a)$, while since $c \in [0, 1]$ and $0 \leq g \leq 1$, we have that $0 \leq g' = cg \leq 1$; hence $g' \in C(\alpha'(a), \nu(\alpha', \mu^*))$. Moreover, obviously $g' \leq g$.

For point (b), let $\hat{\mu}'(a), \hat{\mu}(a)$ be the optimal narratives at $\mu^{*'}$ and μ^* it is sufficient to show that, given any superset $S(r)$ of u it holds $\hat{\mu}'(a)(S(r)) \geq \hat{\mu}(a)(S(r))$. Let ν, ν' be the distribution of outcome under the two models and observe that, by definition of the productivity relation

$$\forall r \quad \nu'(S(r)) \geq \nu(S(r)). \tag{14}$$

[23] Recall that we are treating $\mathrm{P}_{\text{general}}$ as a minimization problem.

Hence, we get that the superset on which $\hat{\mu}'(a)$ concentrates mass is higher than that on which $\hat{\mu}(a)$: $\hat{u}' \geq \hat{u}$. To show our point, note that, given any superset $S(r)$, it holds that

$$\hat{\mu}'(a)(S(r)) = \frac{1}{\alpha(a)}\nu'(S(r) \cap S(\hat{u}')) = \frac{1}{\alpha(a)}\nu'(S(\max\{r, \hat{u}'\})) = \begin{cases} 1 & r \leq \hat{u}' \\ \nu'(S(r))/\alpha(a) & r > \hat{u}'. \end{cases}$$

Analogously,

$$\hat{\mu}(a)(S(r)) = \frac{1}{\alpha(a)}\nu(S(r) \cap S(\hat{u})) = \frac{1}{\alpha(a)}\nu(S(\max\{r, \hat{u}\})) = \begin{cases} 1 & r \leq \hat{u} \\ \nu(S(r))/\alpha(a) & r > \hat{u}. \end{cases}$$

To compare the quantities we distinguish three cases

- $r > \hat{u}' \geq \hat{u}$: $\hat{\mu}'(a)(S(r)) = \nu'(S(r))/\alpha(a) > \nu(S(r))/\alpha(a) = \hat{\mu}(a)(S(r))$ by Equation 14.
- $\hat{u}' \geq r \geq \hat{u}$: $\hat{\mu}'(a)(S(r)) = 1 \geq \nu(S(r))/\alpha(a)$ since $\nu(S(r)) \leq \alpha(a)$ by definition of $\hat{u}$.
- $\hat{u}' \geq \hat{u} > r$ in which case $\hat{\mu}'(a)(S(r)) = 1 = \hat{\mu}(a)(S(r))$.

This concludes our proof.

Proof of Proposition 4 To see 1 note that, for any μ^*, $\delta(\alpha_h, \mu^*)$ is decreasing in α_h, since $V_h(\alpha, \nu(\alpha, \mu^*))$ is decreasing in α_h and $V_l(\alpha, \nu(\alpha, \mu^*))$ is increasing in α_h, as implied by Proposition 2 (note that the h-majorization reduces to the order of reals on $[0, 1]$). Since F_ϕ is increasing, the first point follows.

For point 2 just observe that that when $\alpha_h = \alpha_t = 1/2$, then $\hat{u}^H = \hat{u}^L$, so that narratives $\hat{\mu}^H$ and $\hat{\mu}^L$ attribute the same superset of u (respectively to h and t). Hence, $\delta(1/2) = 0$ and the result follows being the popularity shock symmetric.

For point 3 define $u^* = \max_Y u$ (allowing it to be possibly $+\infty$) let $\underline{\delta} = u^* - \mathbb{E}_{\mu^*(l)}[u(Y)]$ and $\overline{\delta} = \mathbb{E}_{\mu^*(h)}[u(Y)] - u^*$. The result follows noting that $\underline{\delta} = \lim_{\alpha_h \to 0} \delta(\alpha_h)$ and $\overline{\delta} = \lim_{\alpha_h \to 1} \delta(\alpha_h)$.

Proof of Proposition 5 Recall Equation 13. First note that, given any two models $\mu, \mu' \in \mathcal{M}(\alpha, \mu^*)$, subracting Equation 13 evaluated at μ from the same evaluated at μ' we obtain

$$\forall y \quad F_{\mu(l)}(y) - F_{\mu'(l)}(y) = -\left(\frac{\alpha}{1-\alpha}\right)[F_{\mu(h)}(y) - F_{\mu'(h)}(y)].$$

It follows that

$$\begin{aligned}
d^{KS}(\mu(l),\mu'(l)) &= \sup_{y\in\mathcal{Y}} |F_{\mu(l)} - F_{\mu'(l)}(y)| \\
&= \left(\frac{\alpha}{1-\alpha}\right) \sup_{y\in\mathcal{Y}} |F_{\mu(h)}(y) - F_{\mu'(h)}(y)| = \left(\frac{\alpha}{1-\alpha}\right) d^{KS}(\mu(h),\mu'(h)).
\end{aligned}$$

Then, a pair of narratives $(\mu,\mu') \in \mathcal{M}(\alpha,\mu^*)$ maximizes $d^{\mathcal{M}}$ if and only if it maximizes either of the two addends in the definition of $d^{\mathcal{M}}$, for instance $d^{KS}(\mu(h),\mu'(h))$. Then observe that, for any $\mu \in \mathcal{M}(\alpha,\mu^*)$, $d^{KS}(\hat{\mu}^H(h),\mu(h))$ writes, using again Equation 13 constraint and the expression for optimal narratives in Equation 9 we have

$$\begin{aligned}
d^{KS}(\hat{\mu}^H(h),\mu(h)) &= \sup_{y\in\mathcal{Y}} \left| \frac{F_v(y)-(1-\alpha)}{\alpha}\mathbf{1}_{y>\hat{y}^H} - \frac{F_v(y)-(1-\alpha)F_{\mu(l)}(y)}{\alpha} \right| \\
&= \max\left\{ \sup_{y\le\hat{y}^H} \frac{F_v(y)-(1-\alpha)F_{\mu(l)}(y)}{\alpha}, \sup_{y>\hat{y}^H} \frac{(1-\alpha)[1-F_{\mu(l)}(y)]}{\alpha} \right\} \\
&\le \max\left\{ \sup_{y\le\hat{y}^H} \frac{F_v(y)-(1-\alpha)F_{\hat{\mu}^L(t)}(y)}{\alpha}, \sup_{y>\hat{y}^H} \frac{(1-\alpha)[1-F_{\hat{\mu}^L(t)}(y)]}{\alpha} \right\} \\
&= d^{KS}(\hat{\mu}^H(h),\hat{\mu}^L(l)).
\end{aligned}$$

where the last inequality follows from the fact that, $F_{\hat{\mu}^L(t)}(y) \le F_{\mu(l)}(y)$ for any $\mu \in \mathcal{M}(\alpha,\mu^*)$ (i.e. $\hat{\mu}^L(t)$ FOSD $\mu(l)$ for any plausible μ). This is evident from Equation 9, noting that any satisfies $F_{\mu(l)}(y) = \frac{F_v-(1-\alpha(a))F_{\mu(h)}}{\alpha(a)}$ and $F_{\mu(h)} \le 1$.

For the second point, first note that $d(\mu,\mu') \in [0,1]$ for any two models μ,μ'. Then observe that, when $\alpha = 1/2$ it holds that $\hat{y}^H = \hat{y}^L = F_v^{-1}(1/2) \equiv \hat{y}$. Hence we have that, when $\alpha = 1/2$:

$$\begin{aligned}
|F_{\hat{\mu}^H}(\hat{y}|h) - F_{\hat{\mu}^L}(\hat{y}|h)| &= \left| \frac{F_v(\hat{y})-1/2}{1/2} - \frac{F_v(\hat{y})}{1/2} \right| = 1 = \\
&= \sup_y |F_{\hat{\mu}^H}(y|h) - F_{\hat{\mu}^L}(y|h)| = d^{KS}(\hat{\mu}^H(h),\hat{\mu}^L(h)),
\end{aligned}$$

and, similary,

$$\begin{aligned}
|F_{\hat{\mu}^H}(\hat{y}|l) - F_{\hat{\mu}^L}(\hat{y}|l)| &= \left| \frac{F_v(\hat{y})}{1/2} - \frac{F_v(\hat{y})-1/2}{1/2} \right| = 1 = \\
&= \sup_y |F_{\hat{\mu}^H}(y|l) - F_{\hat{\mu}^L(t)}(y|l)| = d^{KS}(\hat{\mu}^H(l),\hat{\mu}^L(l)).
\end{aligned}$$

It follows that $d^{\mathcal{M}}(\hat{\mu}^H,\hat{\mu}^L) = 1$ which is the maximum value for our distance between models.

C For Online Publication – Extensions of the Model

C.1 Connection with Optimal Transport

Intuitively, $\mathrm{P}_{\text{general}}$ consists in correlating optimally two random variables describing, respectively, past policy implementation and outcome realizations. Then, one may wonder whether such a problem is equivalent to that of choosing a particular joint distribution $\gamma \in \Delta(\mathcal{A} \times \mathcal{Y})$ such that its marginals are precisely $marg_{\mathcal{A}}\gamma = \alpha$ and $marg_{\mathcal{Y}} = \nu$. Such distributions are called *couplings* between α and ν and determine a (strict) subset $\Gamma(\alpha, \nu) \subset \Delta(\mathcal{A} \times \mathcal{Y})$. The answer is positive: $\mathcal{M}(\alpha, \nu)$ and $\Gamma(\alpha, \nu)$ can be identified, via the following canonical bijective correspondence[24]

$$\mu \in \mathcal{M}(\alpha,\nu) \mapsto \gamma^{\mu} \quad s.t. \quad \forall A \subseteq \mathcal{A}, Y \subseteq \mathcal{Y} \; meas. \; \gamma^{\mu}(A \times Y) = \int_A \mu(a)(Y) d\alpha(a)$$
$$\gamma \in \Gamma(\alpha,\nu) \mapsto \mu^{\gamma} \quad s.t. \quad \forall a \in \mathcal{A}, Y \subseteq \mathcal{Y} \; meas. \; \mu^{\gamma}(a)(Y) = \frac{\gamma(\{a\} \times Y)}{\alpha(a)}.$$

The correspondence is essentially mapping a plausible narrative into a coupling whose set of conditional distributions is the narrative itself, and viceversa it is mapping a coupling in its set of conditional distributions. Having established this fact, note how $\mathrm{P}_{\text{general}}$ is equivalent to the problem of finding a coupling such that a given conditional distributions maximize the expectation of u. In other words, we can equivalently solve it as

$$\max_{\gamma \in \Gamma(\alpha,\nu)} \int_{\mathcal{Y}} u(y) d\gamma | a(y).$$

Introducing the surplus function, $\Phi_a(a', y) = \frac{u(y)}{\alpha(a)} \mathbf{1}_{a'=a}$we can rewrite

$$\max_{\gamma \in \Gamma(\alpha,\nu)} \int_{\mathcal{A}} \left(\int_{\mathcal{Y}} \Phi_a(a', y) d\gamma | a'(y) \right) d\alpha(a') \quad \text{or equivalently} \quad \max_{\gamma \in \Gamma(\alpha,\nu)} \int_{\mathcal{A} \times \mathcal{Y}} \Phi_a(a', y) d\gamma(a', y). \tag{15}$$

in this form, our problem is indeed an Optimal Transport one.

Moreover, restrict attention to the Monotone Increasing Model. Consider any embedding $\iota : \mathcal{A} \hookrightarrow \mathbb{R}$ such that $\iota(a) = \max \iota(\mathcal{A})$. It is easy to check that, regarding Φ_a as a function on $\mathbb{R}^2$, it is *supermodular*, in the sense that for any $a', a'' \in \iota(\mathcal{A})$ and any $y', y'' \in \mathbb{R}$

$$\Phi_a(\max\{a', a''\}, \max\{y', y''\}) + \Phi_a(\min\{a', a''\}, \min\{y', y''\}) \geq \Phi_a(a', y') + \Phi_a(a'', y'').$$

Indeed, when both a' and a'' are different from a, both the RHS and the LHS equal 0. When

[24] The correspondence does not need α to have full support (an assumption we are mantaining throughout). Indeed, one could just extend the assignment below arbitrarily on a's such that $\alpha(a) = 0$.

both are equal to a, both the RHS and the LHS are equal to $\Phi_a(a, y') + \Phi_a(a, y'')$. When only one of the two is equal to a, say, withou loss a', the inequality reduces to

$$\Phi_a(a', \max\{y', y''\}) \geq \Phi_a(a', y')$$

which holds since u is increasing. Then, applying Theorem 4.3.(i) from Galichon (2018) we obtain that the solution $\hat{\gamma}_a$ to Equation 15 must be the *comonotone coupling* between α and ν. In other words, there exists a uniform random variable $U \sim \mathcal{U}([0,1])$ such that the random vector $(F_\alpha^{-1}(U), F_\nu^{-1}(U))$ is distributed according to $\hat{\gamma}_a$. This alternative way of seeing the optimal narrative formalizes the sense in which the optimal narrative correlates the best ouctomes with intervention a: essentially, such a narrative realizes *positive assortative matching*.

While we do not pursue the optimal transport direction in the current paper, we regard it as an avenue for potential generalziations of our model.

C.2 Alternative Timing for the Probabilistic Voting Model

One potentially concerning aspect of our political competition game is that the voter uses H's narrative to evaluate h and L's narrative to evaluate l, as expressed by Equation 10. A natural alternative is to modify the timing of the game, isolating a narrative choice and a voting stage.

1. In a first stage, the voter chooses to adopt the most hopeful narrative, taking into account that politicians will implement their commitments. In other words, she chooses to believe in μ^H iff

$$\mathbb{E}_{\mu^H(h)}[u(Y)] \geq \mathbb{E}_{\mu^L(l)}[u(Y)] \tag{16}$$

   and in μ^L otherwise. Note that this is same condition expressed by Equation 10, up to the popularity shock.

2. In a second stage, the voter uses the narrative she adopted to evaluate candidates and decides for whom to cast her vote. In particular, let μ^W be the stage one winner and w the corresponding policy. The voter votes for W iff

$$\mathbb{E}_{\mu^W(w)}[u(Y)] \geq \mathbb{E}_{\mu^W(\neg w)}[u(Y)] + \phi \qquad \phi \sim \mathcal{U}\left(\left[-\frac{1}{2\zeta}, \frac{1}{2\zeta}\right]\right)$$

   and for $\neg W$ otherwise, where $W \in \{H, L\}$ and $\neg$ has the obvious meaning.

Note that under this timing it is still optimal for each politician to announce the solution to Problem $\mathrm{P}_{\text{general}}$. But then, it always holds

$$\mathbb{E}_{\hat{\mu}^{W}(\neg w)}[u(Y)] \leq \mathbb{E}_{\hat{\mu}^{\neg W}(\neg w)}[u(Y)],$$

since $\hat{\mu}^{\neg W}$ is the solution to the maximization problem for W[25]. Then, the probability that the first stage winner W wins the election (i.e., in stage 2) in the modified model is at least as large that he wins in the original model, as his "modified narrative advantage" is higher than that defined in Equation 11. Now, fix any coarse memory (α, ν). Without loss, suppose that this is a memory where H has the narrative advantage in the original model (i.e., $\alpha_h \leq \frac{1}{2}$). Then, by Equation 16, $W = H$. Hence, we discover that at memories where H had higher probability of winning than L in the original model, he has even higher probability of winning in the modified one. The same holds for L by symmetry.

Overall, modifying the competition game as above has the effect of making the probability of winning steeper around the fixpoint $1/2$, hence amplifying the effects of narratives. In particular, we expect an increase in the average length of cycles in the dynamic extension of the modified game.

Since the reasoning is conditional on the stage one winner, it extends also to the case where it is stochastic as the result of adding a shock to Equation 16, assuming that the shocks in the two stages are independent.

C.3 Alternative Laws of Motion for Memory

As the intuitive discussion in Section 4 should make clear, the main qualitative feature of our dynamic, namely that plausible narratives produce political cycles, does not rely on the law of motion for memory tracking historical frequencies, but just on the fact that implementation increases the recalled frequency of a policy. To make the point formally, here we consider an arbitrary law of motion of the form

$$\begin{cases} \alpha_h^{\tau+1} & = \kappa^{\tau}\alpha_h^{\tau} + (1-\kappa^{\tau})w^{\tau} \\ \nu^{\tau+1} & = \alpha_h^{\tau+1}\mu^*(h) + (1-\alpha_h^{\tau+1})\mu^*(l). \end{cases}$$

Where (κ^{τ}) is a weakly increasing sequence in $(0,1)$ converging to some $\kappa^{\infty} \in [0,1]$. In other words, sufficient persistency is enough to have cycles. Examples following in this class include both historical averages ($\kappa^{\tau} = \frac{\tau+1}{\tau+2}, \kappa^{\infty} = 1$) and moving averages with fixed persistence ($\kappa^{\tau} = \kappa^{\infty} = \kappa$), which capture limitation in memory length hence, recency bias, analyzed

[25] Actually, the difference will in general be pretty sizeable, since $\hat{\mu}^{W}(\neg w)$ is chosen to *minimize* the same objective

below. Then we can show the following qualitative result on α^τ.

Proposition A6 *If* $\zeta > \max\left\{\frac{1}{2|\underline{\delta}|}, \frac{1}{2||\bar{\delta}|}\right\}$, α_h^τ *is asymptotically bound in*

$$\left[\kappa^\infty \delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa^\infty \delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^\infty)\right] \subset [0,1]. \tag{17}$$

Hence both candidates win infinitely often.

Proof of Proposition A6 Consider any period τ such that $\alpha_h^\tau \in \left(\delta^{-1}\left(\frac{1}{2\zeta}\right), \delta^{-1}\left(-\frac{1}{2\zeta}\right)\right)$. Observe that such a period exists because H wins (loses) for sure in the complement of such an interval. At period $\tau+1$, we will have that $\alpha_h^{\tau+1} \in \left(\kappa^\tau \delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa^\tau \delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^\tau)\right)$. Consider $\alpha^{\tau+2}$, we have three cases:

- If $\alpha_h^{\tau+1} \in \left(\kappa^\tau \delta^{-1}\left(\frac{1}{2\zeta}\right), \delta^{-1}\left(\frac{1}{2\zeta}\right)\right)$, H wins for sure and $\alpha_h^{\tau+2} > \alpha_h^{\tau+1}$. Moreover, $\alpha_h^{\tau+2} < \kappa^{\tau+1}\delta^{-1}\left(\frac{1}{2\zeta}\right) + (1-\kappa^{\tau+1}) < \kappa^\tau \delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^\tau)$ since $\kappa^{\tau+1} \geq \kappa^\tau > \kappa^\tau \frac{1-\delta^{-1}(-1/2\zeta)}{1-\delta^{-1}(1/2\zeta)}$.
- If $\alpha_h^{\tau+1} \in \left(\delta^{-1}\left(\frac{1}{2\zeta}\right), \delta^{-1}\left(-\frac{1}{2\zeta}\right)\right)$ then $\alpha_h^{\tau+2} \in \left(\kappa^{\tau+1}\delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa^{\tau+1}\delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^{\tau+1})\right)$ which is a (weak) subset of $\left(\kappa^\tau \delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa^\tau \delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^\tau)\right)$ since $\kappa^{\tau+1} \geq \kappa^\tau$.
- If $\alpha_h^{\tau+1} \in \left(\delta^{-1}\left(-\frac{1}{2\zeta}\right), \kappa^\tau \delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^\tau)\right)$. T wins for sure and $\alpha_h^{\tau+2} < \alpha_h^{\tau+1}$. Moreover, $\alpha_h^{\tau+2} > \kappa^{\tau+1}\delta^{-1}\left(-\frac{1}{2\zeta}\right) > \kappa^\tau \delta^{-1}\left(\frac{1}{2\zeta}\right)$ since $\kappa^{\tau+1} \geq \kappa^\tau$ and $\left(-\frac{1}{2\zeta}\right) > \left(\frac{1}{2\zeta}\right)$.

Putting all together, we obtain that $\alpha_h^{\tau+2} \in \left(\kappa^\tau \delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa^\tau \delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^\tau)\right)$ and inductively, $\alpha_h^{\tau'} \in \left(\kappa^\tau \delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa^\tau \delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^\tau)\right)$ for any $\tau' > \tau$. The conclusion follows by noting that

$$\left(\kappa^\tau \delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa^\tau \delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^\tau)\right) \xrightarrow{\tau\to\infty} \left[\kappa^\infty \delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa^\infty \delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa^\infty)\right].$$

Limited Memory We can model limited memory considering a law of motion with fixed persistency $\kappa^\tau = \kappa$. In this case, the induced stochastic process for α^τ is a time-homogeneous Markov chain. If at time τ the memory state is α_h^τ, the state at time $\tau+1$ is

$$\alpha^{\tau+1} = \begin{cases} \kappa\alpha_h^\tau + (1-\kappa) & \text{w. prob. } P^H(\alpha_h^\tau) \\ \kappa\alpha_h^\tau & \text{w. prob. } 1 - P^H(\alpha_h^\tau). \end{cases}$$

In this case, clearly the process cannot converge to a number. Then, the asymptotic properties of our system will be governed by the stationary distribution of the chain, which we are able to show exists, is unique, and is supported in the interval 17, for $\kappa^\infty = \kappa$. Since the stationary distribution is ergodic, we can compute the limit of time averages of quantities of interest for our system computing state space average with respect to the stationary distribution. In this sense, we can still study the frequency with which candidates win in the limit, namely the statistic

$$\tilde{W} = \lim_{T\to\infty} \frac{\sum_{t=0}^{T} W_t}{T}.$$

We collect our results in the following statement.

Proposition A7 *The Markov chain for the voter's memory has a unique ergodic stationary distribution $\tilde{\pi} \in \Delta([0,1])$ whose support is included in the interval $\left[\kappa\delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa\delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa)\right]$. Moreover, it holds*

$$\mathbb{P}\left(\tilde{W} = \mathbb{E}_{\tilde{\pi}}[P^H(\alpha)]\right) = 1. \tag{18}$$

The result follows from the application of the (rather technical) theory of discrete-time Markov chains on continuous state spaces[26]. In particular, the second point follows by applying the ergodic theorem to a chain on $[0,1] \times \{0,1\}$ obtained as a lifting of the first, where the second state does not affect transitions, but keeps track of the winner at each period. While it is in general not feasible to characterize analytically the stationary distribution[27] from Proposition A7 we learn that not every state in the interval 17 will be attained with the same frequency asymptotically. In simulations, we discover that the limit of the quantity is still 1/2.

Proof of Proposition A7 Since the state space is uncountable, the chain is described by a transition kernel $\tau : [0,1] \to \Delta([0,1])$, as follows

$$\tau : \alpha \in [0,1] \mapsto P^H(\alpha)\delta_{\kappa\alpha+(1-\kappa)} + (1-P^H(\alpha))\delta_{\kappa\alpha} \in \Delta([0,1]),$$

where $\delta_\bullet$ denotes the Dirac's delta.

We start by verifying a series of properties of τ. Preliminary observe that $[0,1]$ endowed with the usual topology is a Polish space. Then note that:

1. τ is Feller. Consider the action of the kernel τ on any (bounded) continuous function $f \in \mathcal{C}_b([0,1])$, defined by $(\tau f)(x) = \int_{[0,1]} f(y) d\tau(x)$. To show τ is Feller we have to show

[26] Hairer (2021) is also an excellent guide to convergence results.

[27] The only information we have is that the stationary distribution's density $f_{\tilde{\pi}}$ is pinned down by the *detailed balance equations*

$$\forall x \in [0,1] \qquad f_{\tilde{\pi}}(x) = P^H\left(\frac{x-(1-\kappa)}{\kappa}\right) f_{\tilde{\pi}}\left(\frac{x-(1-\kappa)}{\kappa}\right) + \left(1 - P^H\left(\frac{x}{\kappa}\right)\right) f_{\tilde{\pi}}\left(\frac{x}{\kappa}\right),$$

which may be solveable in specific examples.

that τf is a (bounded) continuous function. But this follows directly rewriting

$$(\tau f)(x) = f\left(\frac{x-(1-\kappa)}{\kappa}\right) P^H\left(\frac{x-(1-\kappa)}{\kappa}\right) + f\left(\frac{x}{\kappa}\right) P^H\left(\frac{x}{\kappa}\right),$$

since P^H is bounded and continuous.

2. $(\tau^n(x))_{n\in\mathbb{N}}$[28] is tight for any $x \in [0,1]$. This fact follows trivially from the fact that any positive measure on a Polish space is tight. This fact is known as Ulam's lemma

3. Given any point $x \in I = \left[\kappa\delta^{-1}\left(\frac{1}{2\zeta}\right), \kappa\delta^{-1}\left(-\frac{1}{2\zeta}\right) + (1-\kappa)\right]$, x is acessible for τ. Fix an x in the interval I. We have to show that for every y and every ε there exists k such that $\tau^k(y)((x-\varepsilon, x+\varepsilon)) > 0$. To see this, one can just note that, starting from any point y, the probability distribution $\tau^n(y)$ associated to the n-th iteration of the chain gives positive mass to the points

$$\begin{aligned} S_n = \{\kappa^n y + p(\kappa)(1-\kappa) \mid p(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} \\ for\ (a_0, \ldots, a_{n-1}) \in \{0,1\}^n\} \cap I. \end{aligned}$$

Since the sequence of set (S_n) is dense in I (in the sense that for all $x \in I$ and all ε there exists k such that $S_k \cap (x-\varepsilon, x+\varepsilon) \neq 0$) we can conclude.

4. τ is recurrent. It is succicient to show that there exists a probability measure μ such that for any Borel subset $A \subset [0,1]$ it holds that

$$\mu(A) > 0 \ \Rightarrow\ \forall y \in [0,1]\ \mathbb{P}(\tau_A < \infty | \alpha_0 = y) = 1,$$

where $\tau_A = \inf\{t \geq 0 \mid \alpha_h^\tau \in A\}$. Just consider the uniform probability measure over the interval I. The previous point shows that the implication holds. Indeed, since the chain reaching any ball is a positive probability event, the event $\tau_A = \infty$, which means that A is never reached, has probability 0.

Then, existence of a stationary distribution follows from the following result reported in Hairer (2021)

> **Theorem**: Let τ be a Feller Markov kernel over a Polish space X. Assume that there exists $x_0 \in X$ such that the sequence of measures $(\tau^n(x))_{n\in\mathbb{N}}$ is tight.
> Then, there exists at least one stationary probability measure for τ.

Uniqueness follows from the following result, proved in Bayer et al. (2011)

[28] $\tau^n(x)$ is the measure associated to x by the composition of Markov kernels $\underbrace{\tau \circ \cdots \circ \tau}_{n\ times}$

> **Theorem**: Let τ be a recurrent Markov kernel.
>
> Then, there exists at most one stationary probability measure for τ.

Let us call $\tilde{\pi}$ the unique stationary measure of τ. To see the fact about the support of $\tilde{\pi}$, we can first combine the following result proved in Hairer (2021) with point 3 to show that the $supp(\tilde{\pi}) \subseteq I$

> **Lemma**: Let τ be a Markov kernel over a Poilish space X and let $x \in X$ be acessible.
>
> Then, $x \in supp(\mu)$ for every stationary probability measure.

To conclude about the support, we can directly note that $\tilde{\pi}([0,1] \setminus I) = 0$ since transitions in $X \setminus I$ are deterministic, and hence, if $\tilde{\pi}$ put mass there, we surely would have $supp(\tilde{\pi}) \neq supp\left(\int_{[0,1]} \tau(x) d\tilde{\pi}(x)\right)$.

Ergodicity follows by uniqueness, since any invariant distribution must be written as a covex combination of ergodic invariant distributions[29]

For the second part, first note that the realizations of W_t are determined as follows. One runs the Markov chain (α_h^τ) with the transition probabilities specified by τ and at each period t independently samples a random variable W_t which is Bernoulli with parameter $P^H(\alpha_h^\tau)$. Hence, we can describe the joint evolution of (α_h^τ, W_t) building a Markov chain with state space $[0,1] \times \{0,1\}$. First consider an auxiliary kernel $\varphi : [0,1] \to \Delta(\{0,1\})$ describing the family of Bernoulli's associated to the process. In particular, we have:

$$\varphi : \alpha \in [0,1] \mapsto P^H(\alpha)\delta_1 + (1 - P^H(\alpha))\delta_0 \in \Delta(\{0,1\}).$$

Then, we can build the following transition kernel $\tau_\varphi : [0,1] \times \{0,1\} \to \Delta([0,1] \times \{0,1\})$, mapping:

$$\begin{aligned}(\alpha, w) \in [0,1] \times \{0,1\} \mapsto &\int_{[0,1]} \delta(s) \otimes \varphi(s) d\tau(\alpha)(s) \\ &\equiv P^H(\alpha)\delta_{\kappa\alpha+(1-\kappa)} \otimes \varphi(\kappa\alpha + (1-\kappa)) + (1 - P^H(\alpha))\delta_{\kappa\alpha} \otimes \varphi(\kappa\alpha),\end{aligned}$$

where $\otimes$ denotes the independent product of probability measures. τ_φ is a lifting of τ. It follows that it also admits a stationary ergodic distribution $\tilde{\pi}_\varphi$ which can be obtained from the one of τ as

$$\tilde{\pi}_\varphi = \int_{[0,1]} \delta(s) \otimes \varphi(s) d\tilde{\pi}(s) = \tilde{\pi} \otimes \int_{[0,1]} \phi(s) d\tilde{\pi}(s).$$

[29] We can also check ergodicity deirectly. Indeed, observe that if $K \neq \emptyset$ is an invariant set of τ (so that for all $x \in K$, $supp(\tau(x)) \subseteq K$) we claim that it must be the case that $K \supseteq I$. Then, ergodicity follows by definition since $\tilde{\pi}(K) \geq \tilde{\pi}(I) = 1$. To see the claim we can argue by contradiction. Assume K is invariant, but suppose there is $x \in I \setminus K$. Since I is acessible, there must be a finite transition path path from an element $y \in K$ to any element $x \in I$. But since K is invariant, all the elements of the path, including x must be in K. This concludes the first part of the proof.

Applying the Birkhoff ergodic theorem to the projection $f : [0,1] \times \{0,1\} \to \mathbb{R}$ mapping $(\alpha, w) \mapsto w$ we can conclude. Indeed,

$$\begin{aligned} \lim_{T\to\infty} \frac{\sum_{t=1}^{T} W_t}{T} &= \\ \lim_{T\to\infty} \frac{\sum_{t=1}^{T} f(\alpha_h^{\tau}, W_t)}{T} &= \int_{[0,1]\times\{0,1\}} w d\tilde{\pi}_{\varphi} \\ &= \int_{[0,1]} \left(\int_{\{0,1\}} w d\phi(s) \right) d\tilde{\pi}(s) = \mathbb{E}_{\tilde{\pi}}[P^H(\alpha)], \end{aligned}$$

where in the second last passage we observe f does not depend on α and in the last we use that the expectation of a Bernoulli is its success probability, $\int_{\{0,1\}} w d\phi(s) = P^H(s)$.

D For Online Publication – Details on Suggestive Evidence

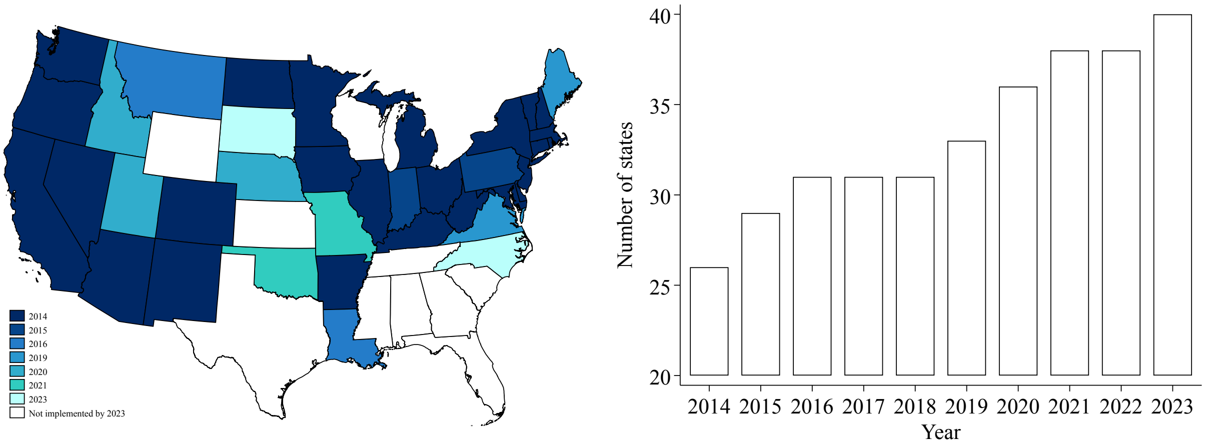
In this Appendix we provide additional details on the suggestive evidence presented at the end of Section 1. First, we briefly document the staggered nature of the Affordable Care Act's implementation across the U.S. states. Then, we discuss in greater detail the findings presented in the main text, focusing on the data employed as well as the procedures adopted.

The major provisions of the Affordable Care Act (ACA) came into force in 2014. However, the decision to implement the Medicaid expansions prescribed by the ACA largely remained at the state level. Consequently, the ACA's implementation has been staggered across different states, as studied by DellaVigna and Kim (2022). Figure A1a represents graphically the staggered implemented across continental states. As clarified further by Figure A1b, only 26 states implemented ACA immediately, while other states gradually followed in the subsequent years. This staggered diffusion of the policy provides a naturally interesting setting for our study, allowing us to utilize a clean event study design. We cannot exclude that the state decision to join the ACA is influenced by factors which also affect our outcomes of interest. Hence, we do not make strong causal claims throughout.

Moreover, the ACA had a desirable, albeit delayed, impact on insurance premiums, at least correlationally. To present evidence of this, we first retrieve data on health insurance premium for all the U.S. states between 2013 and 2019. In particular, we exploit the computation of average monthly premia elaborated by the Heritage Foundation using premium and enrollment data for all individual market plans, which include both ACA-compliant plans and "grandfathered" (pre-ACA) plans, based on data from Centers for Medicare and Medicaid Services. Then, we compute the growth rate of premia with respect to 2013, detrending it for its average growth rate in the United States. In this way, we aim to have a measure of how premia varied differently across the U.S. states. As already discussed, we find that the ACA was associated with a positive and delayed impact on premium's growth rate, as presenting in Figure 2a, where we can clearly see that premia increases less in the states that implemented the ACA. Clearly, while this Figure shows it is reasonable to associate the policy with a desirable movement in premia, it is not enough to claim that such an effect was delayed as it may just be driven by late adopters. For this reason, we present the same graph but split by year of implementation in Figure A2. As we can notice, the largest reduction in the growth of premium is experienced exactly by those states who implemented the ACA earlier, namely in 2014 and 2015, ruling out the previous concern and supporting the idea of a delayed impact.

The second elements of our analysis are the tweets posted by Democrat and Republican congress members elected between 2012 and 2019 and posted in this same time period, and builds on the data employed by Bellodi et al. (2023). This data features a total of about 1.6

Figure A1: Affordable Care Act State Implementation

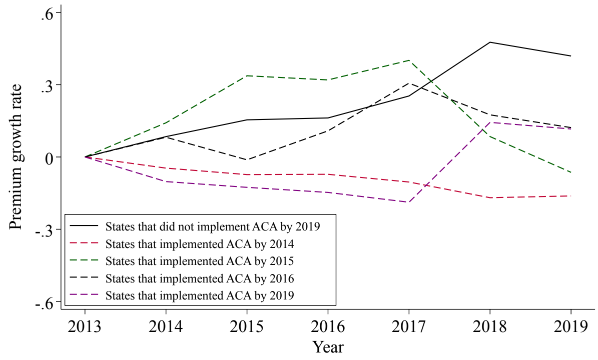


(a) Staggered implementation across U.S. continental states

(b) Cumulative number of states that implemented the ACA by year

Notes: Panel (a) visually describes the adoption year of the Medicaid's expansion prescribed by the ACA for each continental state. Panel (b) shows the cumulative numbers of states that adopted the Medicaid's expansion prescribed by the ACA.

Figure A2: Affordable Care Act and premium by year of implementation.



Notes: The figure presents the growth rate of an aggregate measure of premium detrended by its national growth rate according to the impl

million of tweets, breaking down in 931,446 tweets posted by 496 Democrat congress members and 689,156 tweets posted by 477 Republican congress members.

Our first question is whether Democrats are actually able to take credit for the success of the ACA, and ride the narrative that the ACA is beneficial for the premium shouldered by households. Hence, for this exercise, we are interested in analyzing the narratives spread by politicians about the impact of the ACA on the insurance premium, and how these change before and after the implementation of the policy. However, identifying such narratives in text data and measuring their variation is a non-trivial task, currently at the frontier of text-analysis research. We tackle this issue by focusing on those tweets that talk both about the ACA and the premium. Indeed, despite acknowledging that this technique is not fine-grained, we claim that it still serves our purposes of identifying what we are looking for. Indeed, given that the ACA has been a flagship policy for the Democratic Party, it is unlikely

Table A1: Subsample of tweets about both ACA and premium

Democrats
Great news about 7 of 9 health insurers who participate in the Obamacare market in Michigan reducing their premiums for next year. We'll keep working to make health care and prescription drugs universally affordable. #ForThePeople
Without the ACA's protections for pre-existing conditions, insurance companies will again be able to deny coverage or charge higher premiums for things like high blood pressure, mental illness, or being a woman.
The ACA prevented insurers from raising premiums of Americans with pre-existing conditions. #GrahamCassidy would end that protection.
ACA competition already at work causing insurance premium cuts. Oregon once again leading the way on health care! http://t.co/Z44M7y4dIg
ACA is driving competition, lowering premiums, and protecting families. Can't wait to see more, #GetCovered Oct. 1st! http://t.co/0kjb0ob6AY
Republicans
Statement on today's news of massive health insurance premium hikes in Indiana under Obamacare.#INSen https://t.co/QiHwFbHccu
ObamaCare is causing more premium increases – perhaps as much as 20%. This is not reasonable: http://t.co/lqyYQxoDHK #LASEN
Obamacare = higher premiums for plans Americans don't want or need. #ObamacareRepeal efforts must continue. https://t.co/YfdIclfTG8
ObamaCare has forced higher premiums & lower health care quality. Today I'll vote to partially dismantle the law. https://t.co/1roxu3rVa6
Premiums for Obamacare insurance plans have been rising at a disturbing rate for far too long. Today, I proudly voted to empower #Americans to apply their health insurance subsidy toward the purchase of any qualified #healthplan – on or off the Obamacare exchanges.

that a Democrat would mention both elements claiming that ACA may have the undesirable effect of raising insurance premium. Clearly, the opposite holds for Republicans. Accordingly, as shown by a subsample of the tweets, presented in Table A1, we find that when Democrats (Republicans) mention both terms they do it to stress the positive (negative) effect of ACA on insurance premium. What is more, this is also confirmed by the sentiment analysis[30] we carry out on the tweets about the ACA presented in Figure 2b: Democrats use a more positive tone than Republicans when they talk about the ACA. On a more practical note, we briefly describe the technical procedures employed. First, we remove punctuation, stopwords, numbers, hashtags, Twitter accounts tagged, and links. Then, we stem all the words, to maintain the original meaning of the word by keeping only the root of the words. In turn, we move to identify the tweets talking about the ACA and the premium. For the first task, we mark all the tweet that contain either the roots of the wording "Affordable Care Act", or the root of the word "ACA", or the root of the word "Obamacare". For the second aim, instead, we simply mark all the tweets that contain the root of the word "premium".

Then, exploiting this strategy, we study how the probability a tweet is about premium when it is about ACA varies between Democrats and Republicans with the staggered implementation of the ACA at the state level. That is, we focus on how the implementation of the policy affects the probability that Democrats and Republicans talk about premium in associ-

[30] For the sentiment analysis we rely on the widely used VADER tool (Hutto and Gilbert, 2014), which is particularly suitable for social media content.

ation with ACA. In particular, we carry out an event study at the level of the single politician where we consider the tweets posted in a two-year window around the year of implementation of ACA in the politician's state of election. More precisely, we estimate the following model with year and politician fixed effects separately for Democrats and Republicans

$$
\begin{aligned}
pr_{i,m,t} = \alpha + \beta \times aca_{i,m,t} + \sum_{k=1}^{3} \eta_j (\text{Lag } k)_{i,m,t} + \sum_{k=1}^{3} \lambda_k \times (\text{Lead } k)_{i,m,t} + \varepsilon_{i,m,t} \\
+ \sum_{k=1}^{2} \gamma_k \times (\text{Lag } k)_{i,m,t} \times aca_{i,m,t} + \sum_{k=1}^{2} \delta_k \times (\text{Lead } k)_{i,m,t} \times aca_{i,m,t}. \qquad (19)
\end{aligned}
$$

Where, $pr_{i,m,t}$ is a binary variable taking value of one if tweet i published by congress member m in year t is about premium. Analogously, $aca_{i,m,t}$ takes value of one when it is about ACA Finally, lag and leads describes years elapsed since the year of implementation of the ACA in the state where the politician i has been elected. The results shown in Figure 3a presents the estimates of the γs and the δs separately for Democrats and Republicans, with standard errors clustered at politician level. As already stressed, we find this highly suggestive that Democrats fail to capitalize on the positive outcome associated with the ACA, as, after the implementation of the policy they rather tend to associate less "premium" and "ACA" if compared to Republicans.

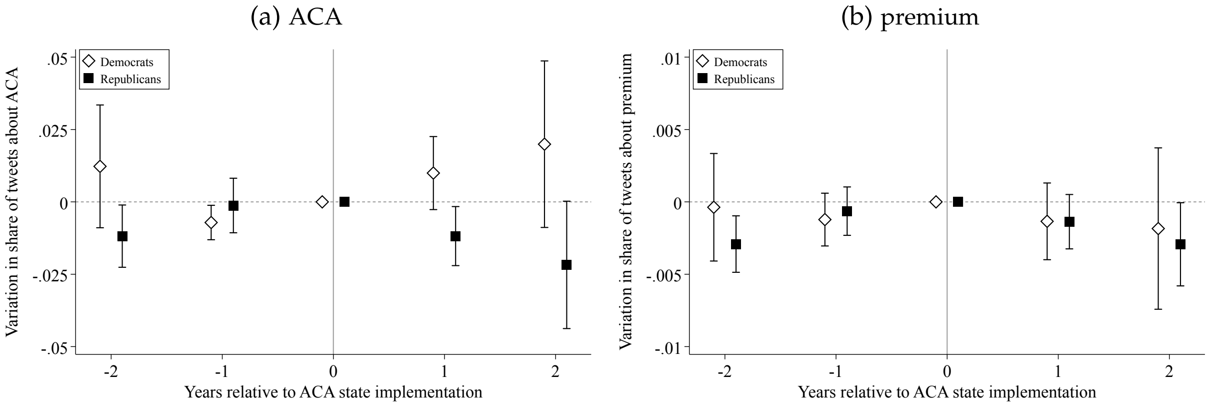
Within this exercise, one may wonder whether Democrats keep talking about the ACA even after its implementation. Indeed, in principle, it could also be that the reduction in narrative is mechanic in the sense that politicians stop talking about a policy once it gets implemented, as they move to the next topic in their agenda. First, we can claim this is unlikely in the case of the ACA. Indeed, even after its implementation, it was a highly debated topic that remained at the center of the U.S. political debates, as the Trump administration fought harshly against it, already in its electoral campaign. To make the point quantitatively, we estimate whether Democrats actually talk less about the ACA after its state implementation. To assess this, we estimate the following model with year and politician fixed effects separately for Democrats and Republicans:

$$
aca_{i,m,t} = \alpha + \sum_{k=1}^{2} \gamma_k \times (\text{Lag } k)_{i,m,t} + \sum_{k=1}^{2} \delta_k \times (\text{Lead } k)_{i,m,t} + \varepsilon_{i,m,t}.
$$

For robustness, we also check whether there is any other mechanic variation in how often Democrats tweet about health insurance *premium*. That is, we also estimate the following model again with year and politician fixed effects for Democrats and Republicans:

$$
pr_{i,m,t} = \alpha + \sum_{k=1}^{2} \gamma_k \times (\text{Lag } k)_{i,m,t} + \sum_{k=1}^{2} \delta_k \times (\text{Lead } k)_{i,m,t} + \varepsilon_{i,m,t}.
$$

Figure A3: Share of tweets about either ACA or premium.



(a) ACA

(b) premium

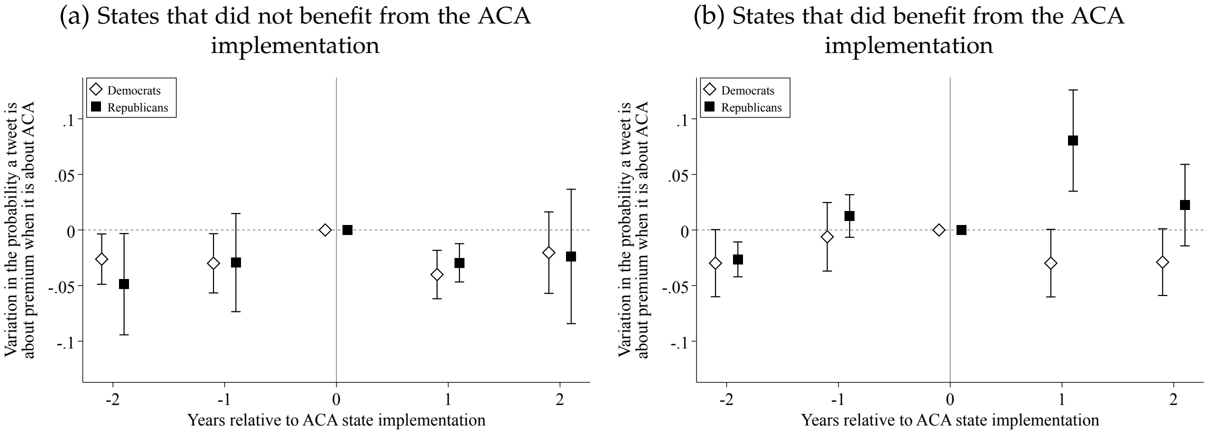
Notes: Panel (a) presents the results of an event study at the politician level describing how the probability a tweet is about ACA varies with the staggered implementation of the Affordable Care Act at the state level within Democrats and within Republicans. Panel (b) presents the results of an event study at the politician level describing how the probability a tweet is about premium varies with the staggered implementation of the Affordable Care Act at the state level within Democrats and within Republicans. In both panels, bars denote 95 percent confidence intervals with standard errors clustered at the politician level.

The estimates, shown in Figures A3a and A3b, makes clear that this concern does not appear to be the case. Indeed, we cannot observe any significant difference in the probability of mentioning *ACA* after its implementation between Democrats and Republicans. If anything, we can notice a slight non-significant increasing trend. On the other hand, concerning the premium, we observe almost no difference in the probability of mentioning it.

What is more, we also mentioned that we find the previous effect to be driven precisely by those state that benefitted the most from the implementation of the ACA (again, at least correlationally). For this aim, we first identify states where, in a two-yer windows around the ACA implementation, the premium decreases after the implementation of the policy. Then, we estimate again model 19 separately for Democrats and Republicans, and also separately for states that did not and did benefit from the implementation of ACA. The results are present in Figure A4a and A4b, respectively. It is possible to notice that, the main finding discussed earlier is driven precisely by states where the ACA was actually associated with a more variation in premium. Indeed, this points exactly to the Democrats' tendency of stopping riding the narrative about the effectiveness of the ACA, despite the observable behavior of premium.

Our second finding points to Democrats and Republicans reducing their disagreement when talking about the ACA after it gets implemented. In this case, the main challenge is to capture semantic similarity among tweets about the ACA. To address it, we rely on well established text-analysis techniques. First, we produce a numerical representation of each tweet using embedding vectors via the pre-trained SBERT language model (Reimers and Gurevych, 2019), which is designed to capture the semantic similarity between sentences. Then, focusing on the tweets about ACA, we compute for each Democrat (Republican) and

Figure A4: Probability a tweet is about premium when it is about ACA for states that did not and did benefit from the ACA implementation



Notes: The figures present the results of an event study at the politician level describing how the probability a tweet is about premium when it is about ACA varies with the staggered implementation of the Affordable Care Act at the state level within Democrats and within Republicans, separetely for states that did not (Panel (a)) and did (Panel (b)) benefit from the ACA in the two year following state implementation. In both panels, bars denote 95 percent confidence intervals with standard errors clustered at the politician level.

for each year the average cosine similarity between their tweets and the tweets posted during the same year by all the Republicans (Democrats) elected in their same state. In this way, we have a measure of how, each year, the tweets about ACA of each politician are semantically similar to those of the opposing politicians, at state level. Then, we assess how this measure of semantic similarity varies with the staggered implementation of the ACA at the state level, carrying out an event study and focusing on a two-year window around the year in which the ACA has been implemented in the state where they have been elected. That is, we estimate different variations of the following model with politician and year fixed effect

$$s_{m,t} = \alpha + \sum_{k=1}^{2} \gamma_k \times (\text{Lag } k)_{m,t} + \sum_{k=1}^{2} \delta_k \times (\text{Lead } k)_{m,t} + \varepsilon_{m,t}.$$

Where $s_{m,t}$ is the average semantic similarity between the tweets about ACA of congress member m posted in year t and those of politician from the opposing party elected in the same state of m in year t. Lag and leads describes years elapsed since the year of implementation of the ACA in the state where the politician i has been elected. The results shown in Figure 3b presents the estimates of the γs and the δs separately for Democrats and Republicans, with standard errors clustered at state level. First, we can observe that before the state implementation of the ACA there are no sizeable differences in the semantic similarity. On the other hand, following the ACA implementation, we observe tweets becoming more similar in their meaning. As anticipated in Section 1, we find suggestive evidence of a reduction in the disagreement following the implementation of the policy. In particular, in the first year after the ACA implementation, we observe a precise estimate describing an increase in the level of similarity with respect to the implementation's year. Altogether, this result supports the prediction of our model: Democrats and Republicans tend to convey more similar messages when talking about the ACA, reducing their disagreement, as the ACA gets implemented.