The Indoctrination Game*

Lotem Ikan† and David Lagziel‡

May 4, 2023

Abstract

The indoctrination game is a complete-information contest over public opinion. The players exert costly effort to manifest their private opinions in public in order to control the discussion, so that the governing opinion is similar to theirs. Our analysis provides a theoretical foundation for the silent majority and vocal minority phenomena, i.e., we show that all moderate opinions remain mute in equilibrium while allowing extremists full control of the discussion. Moreover, we prove that elevated exposure to others' opinions increases the observed polarization among individuals. Using these results, we formulate a new social-learning framework, referred to as *an indoctrination process*.

Journal classification numbers: C72; D62; D72.

Keywords: indoctrination; non-Bayesian social learning; contest theory, polarization.

*For their valuable comments, the authors thank Tomer Blumkin, Ran Eilat, Ehud Lehrer, Igal Milchtaich, Abraham Neyman, Dov Samet, Aner Sela, and Yevgeny Tsodikovich as well as the participants of the Tel Aviv University Game Theory and Mathematical Economics Research Seminar, the 12th annual conference of the Israeli Chapter of the Game Theory Society, and the BIU Game & Economic Theory Seminar.

†Department of Economics, Ben-Gurion University of the Negev, Israel. E-mail: ikanl@post.bgu.ac.il

‡Department of Economics, Ben-Gurion University of the Negev, Israel. E-mail: Davidlag@bgu.ac.il

1 Introduction

The term "belief" is, to some extent, misused in game theory. For example, when someone says "I believe in god", people do not typically assume that this person has a probability distribution (i.e., a belief) on different states of the world that concern the existence of god, and this distribution is to be updated with newly received information. There is neither a distribution, nor informative signals in this context. The same is true for statements such us "I believe in people's right to X". These beliefs are a matter of values and preferences, rather than information. In fact, you do not need to believe in something that you know, and in game theory, a belief actually reflects *knowledge*. That is, a "belief" is the knowledge that certain states may exist, and the knowledge over the probability for each state to be realized.

This distinction is essential because we have a natural tendency to weave together knowledge with perspectives. Consider, for example, the basic structure of what is commonly referred to as a learning model. We begin with some exogenous uncertainty, a randomly chosen state. A rational agent then receives new information, typically a signal, and updates his belief accordingly. This is the fundamental structure of our persuasion models, information-design problems, and Bayesian-learning processes. However, in many real-life scenarios, ranging from politics and economics to religion and sports, this fails to accurately reflect the underlying mechanism. What is the underlying uncertainty in choosing a political side, or in choosing a favorite sports team? In fact, two people can mutually agree on all relevant information, and still disagree on the question of who is the greatest football player of all time. Of course there are some structural uncertainties, but when debating these topics, what we also share are our *opinions*, rather than signals, and opinions work differently. Thus enters the concept of *indoctrination*.

The indoctrination game is a new type of contest in which players hold fixed private opinions that they discuss with others in what could be described as a public debate. The players' main goal is to control the discussion, in the sense that the governing opinion is similar to theirs. More formally, the game comprises a set of individuals whose opinions are distributed on an interval. These individuals exert costly effort to manifest their opinions in public. Their payoffs decrease with the expected distance between their individual opinions and the opinions manifested by others.[1] That is, opposite

[1]The expected distance it taken in absolute value, so that the weights (i.e., the probabilities) are the players' endogenously generated effort levels.

opinions do not offset in the players' payoff functions, and they prefer other individuals, whose opinions are far from theirs, to remain silent. The key ingredient and novelty of this framework is the fact that there is no true state of the world, only different perspectives that collide in equilibrium.

The main goal of this paper is to study the interaction between individuals who hold different opinions, and specifically, the interaction between people who hold moderate opinions and extremists. This goal is divided into three parts. First, we study the equilibria of the indoctrination game given that players completely observe the opinions of others. Next, we analyze a generalized version of the game in which players have limited exposure to others' opinions. Finally, we use the given results to construct a novel social-learning process in which opinions endogenously evolve in future generations. We refer to this dynamic framework as *an indoctrination process*.

To achieve the stated objectives, the paper provides three key results along with several insights. Our first main result, given in Theorem 1, establishes a theoretical foundation for the *silent majority* and the *vocal minority* phenomena. It shows that moderate opinions remain mute in equilibrium (i.e., the silent majority), while giving extremists the ability to govern the discussion. Moreover, our analysis indicates that the individuals' inclination to manifest their opinions is inversely related to their level of representation (i.e., the vocal minority). This negative relation is two-dimensional, depending on the distance between the opinions of the extreme groups, and their sizes. In other words, once extreme groups reduce in size, or become more extreme, the actions of every remaining individual in these groups intensify on average. These phenomena were empirically documented by Mustafaraj et al. (2011) in the context of political discussions on social media. Their findings were recently supported by a Pew Research Center report, entitled "National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Tweets", which states that 97% of political tweets of U.S. adults originate from just 10% of the users, who also hold the least moderate views over the opposing political side.

The intuition behind this crowding-out effect traces back to the augmented relative impact of extreme individuals, one over the other, compared to their impact on moderate players. Extremists typically try to mitigate the effect of the opposing side, so they naturally exert a higher level of effort on aggregate. This aggressive behaviour dilutes the impact of moderate players, thus creating a positive feedback loop that intensifies the extremists' behaviour. The effect eventually stabilizes once all moderate players withdraw from the debate. This result holds independently of the number of players and opinions.

The second main result relates to the extended model in which players only have partial monitoring

over others. In this set-up, we study how the exposure level of individuals to others' opinions affects the equilibria of the game. Our analysis shows that an elevated exposure level increases polarization. To see this, we adapt the seminal polarization metric of Esteban and Ray (1994) to our setting, and show that polarization increases in any equilibrium, as a function of the exposure level. Interestingly, this phenomenon was also empirically documented in a recent field experiment by Bail et al. (2018), who use politically leaning bots on social media to show that exposure to opposing views increases political polarization.

Focusing on the first two parts of the paper, it is clear that the act of indoctrination, within the given framework, is rather futile. The players' main objective, as indicated by their payoff functions, is to influence others by controlling the discourse. However, in this one-stage setting, players do not alter their opinions. This issue is addressed in the third part of this paper, which delves into a new evolutionary, adaptive-learning framework.

In the third part of this paper, we use the equilibrium result of the limited-exposure model to *endogenously* generate a transition matrix between opinions that yields an inter-generational adaptive-learning process. Specifically, in every stage, individuals act according to some equilibrium profile, and the distribution of opinions of the subsequent generation is determined in proportion to the observed opinions given that profile. This generates a non-Bayesian evolutionary process where the transition matrix is repeatedly derived according to the newly realized equilibrium. Our analysis focuses on the stationary distribution of the learning process as a function of the exposure level, and shows that a higher exposure level leads to a more polarized society. In other words, we demonstrate that the distribution of opinions spreads further apart as the exposure level increases.[2]

1.1 Relation to literature

Our basic framework lays in the vast literature of contests which goes back to the seminal study of Tullock (1980), and later followed by Skaperdas (1996) and Baye and Hoppe (2003), among many others. Within this set of games, there exists a specific class of *contests with externalities*,[3] motivated by the early work of Buchanan (1980), and more substantially by the later work of Congleton (1989)

[2]Note that in this context, *"social learning"* refers to the cognitive and evolutionary process of observing and absorbing the subjective opinions of others and should not be confused with either Bayesian or other forms of rational learning.

[3]This feature ranges also to auction theory, which accommodates a vast literature on identity-dependent externalities; see, for example, Funk (1996), Jehiel et al. (1996, 1999), Varma (2002), Aseff and Chade (2008), and Brocas (2013).

which studies status-seeking contests with externalities that affect outside (non-strategic) individuals. In recent years, this research area expended in various directions,[4] thus we shall focus on studies that are closest to the current research agenda.

The early studies of Tullock contests were generalized by Linster (1993), that derives an equilibrium in pure strategies in a setting where losing players are not indifferent to the identity of the winner. Although our framework focuses on different payoff functions, the linear cost function allows us to use similar mathematical methods as the ones used by Linster (1993). Another key feature of our setup goes back to the work of Nitzan (1991), which studies Tullock contests where players are partitioned into groups who compete together. Once a group wins the prize, they apply various sharing rules to divide the prize among its members. The concept of partitioning players into competing groups is quite natural in the context of public debates, and would indeed prove important in our setup, as well.

Similarly to Moldovanu et al. (2012) and Sela (2020), the indoctrination game also encompasses negative externalities. In our framework however all payoffs are negative, rather than a combination of prizes and penalties, carrots and sticks. In general, contests with externalities could also be classified according to the type of externalities and the individuals that are affected by them. The indoctrination game falls within the set of contests with negative, identity-dependent externalities that affect all players, independently of their winning status.

Overall, the study that is closest to the first two parts of this paper is the seminal work of Esteban and Ray (1999), and specifically Section 5 therein. Our basic model extends Esteban and Ray (1999), by generalizing the payoffs and groups of players (using the "linear alienation" given in Esteban and Ray, 1994) and by focusing on different cost functions (similarly to Linster, 1993). Evidently our results give rise to additional conclusions, the obvious one being that the silent-majority and vocal minority phenomena, and the emergence of the stated crowding-out effect in equilibrium.

The third part of our study, which builds on the first two parts, lays in the intersection of adaptive learning and evolutionary processes. Our analysis, however, is closer to the former rather than the latter. The concept of adaptive learning can be traced back to the work of DeGroot (1974), which investigates the stochastic process of reaching a consensus by adapting observed opinions. Our motivation and general objective are quite similar to this line of research. In many of these studies,

[4]See, e.g., Chung (1996), Lee and Hyeong Kang (1998), Eggert and Kolmar (2006), Shaffer (2006), Konrad (2006), Lee (2007), Cohen et al. (2008), Chowdhury and Sheremeta (2011), Ahn et al. (2011), Klose and Kovenock (2015), and Park and Lee (2019), among many others.

players follow certain heuristics, such as Naïve learning as in Golub and Jackson (2010) and Amir et al. (2021), or the majority dynamics as in Galam (2002) and Arieli et al. (2023), that do not necessarily establish an equilibrium in the relevant framework. Our analysis offers a different perspective in two fundamental issues: first, we base the learning process on the equilibria of the limited-exposure indoctrination game; and second, there is no true state of the world in our setting, only opinions. Notably, this allows us to link contest theory with social learning while providing a micro-founded framework for non-Bayesian adaptive processes.

2 The game

The indoctrination game is a complete-information, single-stage contest in which players hold fixed individual opinions that they manifest in public. To do so, the players exert costly effort and are being rewarded according to the distance between the aggregate distribution of publicly observed opinions and their private ones. In equilibrium, players balance their individual cost of effort with the need to shift the public opinion towards their own.

Formally, fix $k \geq 2$ distinct values $O_1 < O_2 < \cdots < O_k$ in $\mathbb{R}$, that represent k different opinions. We shall refer to O_1 and O_k as the *extreme opinions*, and to all others as *moderate* ones.[5] Let $N = \{1, 2, \ldots, n\}$ be the set of players, and for every $i = 1, \ldots, k$, let N_i denote the non-empty set of players with a private opinion O_i, such that $n_i = |N_i| \geq 1$ and $n = \sum_i n_i$. We refer to the players in N_i as the O_i-players.

The action set of every player is $\mathbb{R}_+$. An action $e_j \geq 0$ is the effort that player $j \in N_i$ exerts to publicly manifest his opinion O_i. Given a non-zero action profile $\mathbf{e} = (e_1, \ldots, e_n) \in \mathbb{R}^n_+$, consider the random variable $X_{\mathbf{e}}$ distributed according to

$$\Pr(X_{\mathbf{e}} = O_i) = \frac{\sum_{j \in N_i} e_j}{\sum_{j=1}^{n} e_j} = \frac{E_i}{\sum_{j=1}^{k} E_j},$$

where $E_i = \sum_{j \in N_i} e_j$ is the sum of efforts of all O_i-players. Intuitively, $P_{X_{\mathbf{e}}}(\cdot)$ is the distribution of publicly observed opinions, weighted according to the players' effort levels. If, for example, all O_i-players exert relatively high effort levels (on aggregate and compared to all other players combined), then their opinion would dominate the debate and $X_{\mathbf{e}}$ would be distributed accordingly.

[5] To facilitate the exposition, we sometimes relate to players with extreme/moderate opinions as extreme/moderate players, respectively.

The expected payoff of player $j \in N_i$, given a non-zero effort profile $\mathbf{e} \in \mathbb{R}^n_+$, is

$$U_j(\mathbf{e}|O_i) = -e_j - \mathbb{E}[|O_i - X_{\mathbf{e}}|].$$

The payoff function presents the classic tension in contest theory between the private cost of effort e_j and the need to govern the debate. The term $\mathbb{E}[|O_i - X_{\mathbf{e}}|]$ is the expected distance between opinion O_i and publicly observed opinions, given the players' effort levels $\mathbf{e}$. Thus, in case the distribution of publicly observed opinions $X_{\mathbf{e}}$ shifts towards O_i, then all O_i-players benefit from the reduced expected distance $\mathbb{E}[|O_i - X_{\mathbf{e}}|]$. Note that the expected distance is taken in absolute value, so opposing opinions (relative to O_i) do not offset. To eliminate trivial results of a null debate in which no player exerts positive effort (i.e., to exclude $e_0 = (0, 0, \ldots, 0)$ as an equilibrium), fix $U_j(e_0|O_i) = \inf_{\mathbf{e} \in \mathbb{R}^n_+ \setminus \{e_0\}} U_j(\mathbf{e}|O_i)$ for every opinion O_i and for every player j.[6]

3 The silent majority and the vocal minority

Our analysis begins with equilibria characterization. Theorem 1 describes the equilibria of the indoctrination game, and doing so, presents two intriguing phenomena. The first, referred to as the *silent majority*, shows that all moderate players, i.e., players who do not posses extreme opinions, remain silent in every equilibrium. The theorem formally states that, in every equilibrium, the effort level of every moderate individual is zero. In other words, the only players who extract positive effort levels in equilibrium are the ones who hold the extreme opinions O_1 and O_k.[7]

The second phenomenon, which complements the first, is referred to as the *vocal minority*. Not only that the extreme opinions completely govern the public debate, the average expected effort of every individual in these groups is inversely related to theirs sizes. In other words, individuals from smaller extreme groups tend to be louder on average. This follows from the fact that the aggregate effort of each of these groups in equilibrium depends solely on the distance $|O_1 - O_k|$. So if one group is smaller than the other, the average "vocality" (i.e., effort level) of every individual in the smaller group is higher. Before presenting Theorem 1, we emphasize that the results are independent of the relative position of opinions and the number of moderate players. This underscores the robust nature

[6] Nash equilibria are robust to affine payoff transformations, so if needed, one can adjust the payoff functions to get strictly positive payoffs under undominated strategies.

[7] We acknowledge that the majority could be based in the extremes. The terminology relates to the typical case in which the extremists are relatively small groups.

of the two aforementioned phenomena.

Theorem 1. *In every equilibrium, the effort level of every moderate player is zero, whereas the aggregate effort levels of all extreme players are* $E_1 = E_k = \frac{|O_1 - O_k|}{4}$.

An immediate corollary, following Theorem 1, relates to the unique symmetric equilibrium in which every extreme player exerts the same level of effort as all other players sharing the same opinion. (The proof is follows immediately from Theorem 1, thus omitted.)

Corollary 1. *There exists a unique symmetric equilibrium* $\mathbf{e}^{\text{sym}}$ *such that*

$$\mathbf{e}_j^{\text{sym}} = \begin{cases} 0, & \forall j \in N_i, i \neq 1, k, \\ \frac{|O_1 - O_k|}{4n_i}, & \forall j \in N_i, i = 1, k, \end{cases}$$

and the expected payoff of every player j*, given* $\mathbf{e}^{\text{sym}}$*, is*

$$U_j(\mathbf{e}^{\text{sym}}|O_i) = -\frac{|O_1 - O_k|}{2} \cdot \left[1 + \frac{1}{2n_i} \mathbb{1}_{\{i=1,k\}}\right].$$

The driving force and intuition behind this result is the *crowding-out* effect of extreme players over moderate ones in equilibrium. The impact of extreme players from both sides, one over the other, is significantly higher than their impact on moderate players (in proportion to the distance between the different opinions). So extreme individuals naturally aim to mitigate the effect of other extreme players by increasing their effort levels. This joint "aggressive" behaviour dilutes all other opinions (note that the denominator in $P_{X_{\mathbf{e}}}(\cdot)$ becomes larger), so individuals with moderate opinions are less inclined to extract effort, thus producing a positive feedback loop which results in the stated equilibrium. This is a somewhat extensive, yet natural, *crowding-out* effect in equilibrium. The effect stabilizes once all moderate opinions withdraw from the public debate, whereas the aggregate effort levels of the extreme individuals adjust to $\frac{1}{4}|O_i - O_k|$.

There are several additional conclusions that one can derive from Theorem 1: (i) The crowding-out effect is beneficial for moderate players who retain a strictly higher expected payoff, compared to extremists. Moderate individuals actually increase their payoff by not participating in the public debate, whereas extreme players are bound to invest heavily in this contest; (ii) Everyone lose from polarization. The expected payoffs of all players are proportional to $|O_1 - O_k|$, so additional separation between extreme opinions is detriment. Moreover, extreme players lose the most from polarization; (iii) Free-riding may originate in equilibrium within each group of extreme players. The aggregate

effort levels of extreme individuals are independent of the groups' sizes, so extreme players benefit from the participation of others extremists within the same group. This is supported by Corollary 1 which shows that, under the unique symmetric equilibrium, the expected payoff of extreme players increases with their groups' sizes; and (iv) The equilibria of the game are independent of the relative position and of the number of moderate players. In other words, the relative position of the polarized groups is the key factor that "sets the tone" in the debate. Yet, we stress that this result may change if we divert from a linear cost function, specifically to either convex, or concave cost functions.

4 Limited exposure in public debates

The basic indoctrination game builds on the premise of full monitoring, i.e., that individuals fully observe the opinions of all others. In practice, however, the exposure and attention of players vary, so one should also consider the possibility of a partial-monitoring setting in which players have limited exposure to others' opinions. These limitations could arise from external reasons such as network effects, as well as internal ones, e.g., to preempt cognitive inconsistencies. Namely, when people only partially agree with some ideas, they may refrain from spreading them, thus affecting the ability of others to observe these ideas. Moreover, even if some opinions eventually do become public, people may feel an internal urge to partially ignore them, specifically because they do not match their private ones.

In this section we study how limited exposure/attention to opposing views impacts visible polarization in the debate.[8] Our results show that, at least in the short term (i.e., as long as opinions do not shift), an elevated exposure to opposing opinions has an adverse effect on polarization, making the debate more intense. To gain some preliminary intuition for this statement, consider splitting the basic indoctrination game (given in Section 2) into two separate games, each with at least two opinions, so that the first contains all players with opinions O_1 through $O_{\lfloor k/2 \rfloor}$, and the second contains all players with opinions $O_{\lfloor k/2 \rfloor+1}$ through O_k. Theorem 1 predicts that the extreme individuals within each of these sub-games would control the discussion in proportion to $|O_1 - O_{\lfloor k/2 \rfloor}|$ and $|O_{\lfloor k/2 \rfloor+1} - O_k|$, respectively. In other words, the fragmentation into two separate sub-games reduces the (internal) intensity within each debate. Thus, the reverse procedure through which distinct sub-groups better observe

[8] To simplify the exposition, we follow the limited-exposure terminology in this section, but one could similarly interpret all results to limited attention.

each other, evidently generates a high-intensity debate in equilibrium. This provides some intuition for the conclusion that the debate intensifies the more people are exposed to others' opinions, and it also provides a conceptual framework for the recent empirical evidence provided by Bail et al. (2018) who show how exposing people to opposing views in social media increases political polarization. To formally discuss and prove these statements, we first define a *limited-exposure* indoctrination game, and then adjust the polarization metric of Esteban and Ray (1994) to our context.

To capture the notion of partial monitoring, we introduce an *exposure level* $\delta \in (0,1]$ which limits the ability of players to observe distant opinions. More formally, consider the previously defined indoctrination game, but assume that a fraction of the information that a O_l-player generates is discarded, by a factor of $\delta^{|i-l|}$, until it reaches a O_i-player. In such a case, the payoff function of every player $j \in N_i$ takes the following form

$$U_j(\mathbf{e}|O_i) \quad = \quad -e_j - \frac{\sum_{l=1}^{k} E_l \delta^{|O_i - O_l|} |O_l - O_i|}{\sum_{l=1}^{k} \delta^{|O_i - O_l|} E_l}.$$

In simple terms, the players' exposure to each other decreases as a function of the distance between their individual opinions.

Remark 1. *Before we elaborate on the polarization metric, let us clarify that the analysis throughout this section is confined to a symmetric set-up with three opinions, i.e., $k = 3$ and $|O_1 - O_2| = |O_2 - O_3| = 1$. This assumption is imposed for tractability, and the analysis of the general case, with any number of opinions and valuations, is left for future research. We refer to this limited framework as the* limited-exposure *indoctrination game.*

To measure polarization in public debates, we follow the seminal work of Esteban and Ray (1994) that axiomatically construct the following polarization metric for populations with various characteristics (see Theorem 1 and 2, as well as Section 5.1, therein). We adopt their metric by taking E_i to be the observed volume of opinion i, so that the effort profile $\mathbf{e} \in \mathbb{R}^n_+$ translates to a polarization value of

$$P(\mathbf{e}) = \frac{\sum_{i,j} E_i^2 E_j |O_i - O_j|}{[\sum_i E_i]^3}. \tag{1}$$

This polarization metric is invariant to the aggregate volume of opinions, and typically increases once masses are shifted towards the extremes (see Axioms $1-3$ and Condition H in Esteban and Ray, 1994). Notably, the result given in Theorem 1 above supports the highest possible level of polarization in the general case (of k opinions).[9]

[9] See Theorem 2 in Esteban and Ray (1994) concerning the bimodal distribution.

The polarization level $P(\mathbf{e})$ clearly depends on the induced profile $\mathbf{e} \in \mathbb{R}^n_+$ in equilibrium, which in turn depends on the exposure level δ. So, any discussion about polarization must first specify the equilibrium profile $\mathbf{e}$. For this purpose, we take the broad objective of considering the impact of the exposure level on *all* possible equilibria. Formally,

Definition 1. *let $\Lambda(\delta)$ be the set of all equilibria in the limited-exposure indoctrination game with exposure level δ. We say that the polarization in the limited-exposure indoctrination game* increases *in its exposure level if $P(\mathbf{e}_1) > P(\mathbf{e}_2)$, for every $\mathbf{e}_1 \in \Lambda(\delta_1)$, every $\mathbf{e}_2 \in \Lambda(\delta_2)$, and every $\delta_1 > \delta_2$.*

In other words, we do not restrict our analysis through some equilibrium selection, but consider all possible equilibria of the limited-exposure game.

Our main result in this section, given in Theorem 2 below, indeed shows that the polarization in a given game increases in its exposure level. The intuition behind this result is the augmented relative impact of extreme players on each other, relative to their impact on moderate players. Once the exposure increases, the relative impression of extreme players on each other becomes significant, so that they manifest their opinions more strongly, thus diluting the impression of all moderate players and making the polarization evident.

Theorem 2. *The polarization level of the limited-exposure game strictly increases in its exposure level.*

To prove Theorem 2 we require the following supporting lemma which states that, in any equilibrium, moderate players become relatively less vocal once the exposure increases.

Lemma 1. *For any given exposure level, the ratio between the aggregate effort level of moderate players and that of extreme players, in every equilibrium, is unique and strictly decreases in δ.*

Figure 1 depicts the functional relation, described in Lemma 1, between $\frac{E_2}{E_1+E_3}$ and the exposure level δ in any equilibrium of the limited-exposure game. The relation is implicitly given by the following equation

$$4\left(\delta + \tfrac{E_2}{E_1+E_3}\right)^3 = \left[1 + \delta^2 + 2\delta\tfrac{E_2}{E_1+E_3}\right]^2,$$

as derived in the proof of Lemma 1. In case δ tends to 1, one can see that we converge to the baseline model studied in the previous section, so that the ratio $\frac{E_2}{E_1+E_3}$ tends to zero in equilibrium.

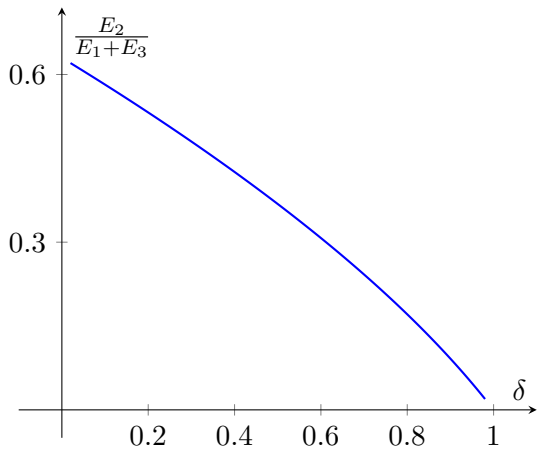

Figure 1: The ratio between the aggregate effort levels of moderate players to that of extreme players, in equilibrium, as a function of the exposure level. Though the equilibrium is not unique, the relation between $\frac{E_2}{E_1+E_3}$ and δ does hold in every equilibria of the limited-exposure 3-player indoctrination game.

5 Dynamic opinions: an indoctrination process

The limited-exposure indoctrination game allows us to discuss, at least in general terms, the possibility of dynamic opinions. Consider, for example, the basic majority-rule (reaction-diffusion) model as in Galam (2002), in which people are repeatedly and randomly matched into subgroups, so that in every stage, each individual adapts the opinion of the majority within his group. To some extent, this is a reduced-form non-strategic model of indoctrination, in which people simply conform to the opinions of others.

To extend this model to our strategic setting, we propose an updated framework called the *indoctrination process*, which involves two adjustments. First, instead of assuming a fixed set of individuals, we consider an inter-generational process where players are replaced in every stage. Second, instead of using the majority rule, we determine the opinions of the newly formed players in each stage based on the distribution of opinions and equilibrium profile from the previous stage.

More formally, for every $\delta \in (0,1]$ and for every stage $t \geq 0$, denote by $\mathbf{e}^t$ an equilibrium profile of the limited-exposure game (assuming that all opinions are represented), and consider the 3×3 transition matrix Q^t with entries $Q^t_{i,j} = \Pr(X_{\mathbf{e}^t} = O_j | O_i)$. Explicitly,

$$Q^t = \begin{bmatrix} \frac{E_1}{E_1+\delta E_2+\delta^2 E_3} & \frac{\delta E_2}{E_1+\delta E_2+\delta^2 E_3} & \frac{\delta^2 E_3}{E_1+\delta E_2+\delta^2 E_3} \\ \frac{\delta E_1}{\delta E_1+E_2+\delta E_3} & \frac{E_2}{\delta E_1+E_2+\delta E_3} & \frac{\delta E_3}{\delta E_1+E_2+\delta E_3} \\ \frac{\delta^2 E_1}{\delta^2 E_1+\delta E_2+E_3} & \frac{\delta E_2}{\delta^2 E_1+\delta E_2+E_3} & \frac{E_3}{\delta^2 E_1+\delta E_2+E_3} \end{bmatrix}.$$

We use this matrix structure to define the following dynamic process. In stage $t = 0$, the players' opinions are fixed according to some initial distribution π_0 with full support. These players act according to an equilibrium profile $\mathbf{e}^0$. In stage $t = 1$, a new generation is formed, and their opinions are distributed according to $\pi_1 = \pi_0 Q^0$, where Q^0 is the previously defined transition matrix associated with $\mathbf{e}^0$. In simple terms, the generation in stage $t = 1$ observes the public opinion generated by the previous generation, which depends both on the equilibrium profile $\mathbf{e}^0$ and on the initial distribution π_0. Subsequently, in each stage $t \geq 1$, the newly formed generation adapts the opinion distribution π_t according to the following equation: $\pi_t = \pi_{t-1} Q^{t-1}$, where Q^{t-1} is the transition matrix associated with the equilibrium $\mathbf{e}^{t-1}$. This process continues indefinitely.[10]

The indoctrination process works such that a newly formed generation observes the opinions of the previous one in equilibrium, while taking into account the different perspectives of each subgroup. This process builds on an inherent biased, as the previous distribution of opinions can significantly influence the subsequent one through the observed opinions. For instance, if the newly formed generation belongs to a population that is heavily skewed in favor of a particular opinion, say O_1, then their opinions would be significantly influenced by the viewpoints of O_1-players in equilibrium. Now, we can use the generic equilibrium profile given in the proof of Lemma 1 to explicitly present the transition matrix in every stage t.

Observation 1. *The transition matrix in every stage t and in every equilibrium $\mathbf{e}^t$ (as given in the proof of* Lemma 1*) is*

$$Q^t = \begin{bmatrix} \frac{1}{1+\delta r^*+\delta^2} & \frac{\delta r^*}{1+\delta r^*+\delta^2} & \frac{\delta^2}{1+\delta r^*+\delta^2} \\ \frac{\delta}{2\delta+r^*} & \frac{r^*}{2\delta+r^*} & \frac{\delta}{2\delta+r^*} \\ \frac{\delta^2}{1+\delta r^*+\delta^2} & \frac{\delta r^*}{1+\delta r^*+\delta^2} & \frac{1}{1+\delta r^*+\delta^2} \end{bmatrix},$$

where $r^* = \frac{E_2}{E_1}$.

Note that this is a right centrosymmetric transition matrix, i.e., it is symmetric with respect to its center $Q^t_{2,2}$ and every row sums to one. Moreover, as long as all opinions are represented, the ratio $r^* = \frac{E_2}{E_1}$ is independent of the number of players holding each opinion. So, for every $\delta \in (0, 1)$, this irreducible and aperiodic transition matrix holds in every stage t and in every equilibrium $\mathbf{e}^t$. Thus,

[10] If π_t contains irrational values, it will not be feasible to implement it with a finite set of players. In such cases, one can use a sufficiently close approximation of π_t, which would also yield sufficiently close results. The notion of M-absorbing sets, as discussed in Lehrer and Shaiderman (2021), is helpful in this regard.

the convergence towards its unique, stationary, probability eigenvector π is guaranteed independently of the initial distribution of opinions. Specifically, its stationary distribution is

$$\pi = \left(\frac{\sqrt{\delta + \frac{1}{2}r^*}}{2\sqrt{\delta + \frac{1}{2}r^*} + r^*}, \frac{r^*}{2\sqrt{\delta + \frac{1}{2}r^*} + r^*}, \frac{\sqrt{\delta + \frac{1}{2}r^*}}{2\sqrt{\delta + \frac{1}{2}r^*} + r^*} \right).$$

Lemma 1 states that r^* is a decreasing function of δ, so one can easily prove that π_2 is decreasing in δ as well, thus establishing that the population becomes more polarized as δ increases.

Lemma 2. *The proportion π_2 of moderate players decreases in $\delta \in (0, 1]$.*

Besides monotonicity, we can use the functional relation between δ and r^*, given after Lemma 1, to compute the stationary distribution in case δ tends to either 0 or 1. Specifically, in case δ tends to 0, the stationary distribution converges to $\pi = \left(\frac{1}{2+2^{1/3}}, \frac{2^{1/3}}{2+2^{1/3}}, \frac{1}{2+2^{1/3}} \right) \approx (0.307, 0.386, 0.307)$. On the other hand, in case δ tends to 1, we know that r^* converges to 0, and so we get $\pi = (0.5, 0, 0.5)$ in case of full exposure. In other words, if there are no limitations and everyone can observe all opinions, the entire population reaches the most extreme state of polarization.

6 In conclusion

The indoctrination game offers a valuable perspective on social debates, which goes beyond the formal results presented in this paper. It presents an alternative framework to the standard Bayesian inference and rational-learning models, allowing for players to indoctrinate each other. This shift in perspective prompts a reevaluation of the assumption that there is always an objective, unknown state of the world that individuals seek to discover. Instead, it recognizes the possibility that people may hold differing opinions based on their subjective life experiences. The game provides a theoretical foundation for empirically documented phenomena such as the silent majority and vocal minority, as well as the impact of exposure to opposing opinions on polarization within a population. However, the game should not be regarded as a restrictive approach to social debates, but rather as an alternative framework that allows for a more nuanced understanding of how people form and revise their beliefs in social settings.

References

Ahn, T. K., R. Mark Isaac, and Timothy C. Salmon, "Rent seeking in groups," *International Journal of Industrial Organization*, January 2011, *29* (1), 116–125.

Amir, Gideon, Itai Arieli, Galit Ashkenazi-Golan, and Ron Peretz, "Granular DeGroot Dynamics – a Model for Robust Naive Learning in Social Networks," February 2021.

Arieli, Arieli, Galit Ashkenazi-Golan, Ron Peretz, and Yevgeny Tsodikovich, "Minimal contagious sets in innovation diffusion networks," 2023.

Aseff, Jorge and Hector Chade, "An Optimal Auction with Identity-Dependent Externalities," *The RAND Journal of Economics*, 2008, *39* (3), 731–746.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky, "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National Academy of Sciences*, September 2018, *115* (37), 9216–9221.

Baye, Michael R. and Heidrun C. Hoppe, "The strategic equivalence of rent-seeking, innovation, and patent-race games," *Games and Economic Behavior*, August 2003, *44* (2), 217–226.

Brocas, Isabelle, "Optimal allocation mechanisms with type-dependent negative externalities," *Theory and Decision*, September 2013, *75* (3), 359–387.

Buchanan, James M, "Rent seeking under external diseconomies," in "Buchanan, J.M., Tollison, R.D., and Tullock, G., (Eds.) , Toward a Theory of the Rent-seeking Society" number 4. In '1.', College Station: Texas A&M University, 1980, pp. 183–194.

Chowdhury, Subhasish M. and Roman M. Sheremeta, "A generalized Tullock contest," *Public Choice*, June 2011, *147* (3), 413–420.

Chung, Tai-Yeong, "Rent-Seeking Contest When the Prize Increases with Aggregate Efforts," *Public Choice*, 1996, *87* (1/2), 55–66. Publisher: Springer.

Cohen, Chen, Todd R. Kaplan, and Aner Sela, "Optimal Rewards in Contests," *The RAND Journal of Economics*, 2008, *39* (2), 434–451.

Congleton, Roger D., "Efficient status seeking: Externalities, and the evolution of status games," *Journal of Economic Behavior & Organization*, March 1989, *11* (2), 175–190.

das Varma, Gopal, "Standard Auctions with Identity-Dependent Externalities," *The RAND Journal of Economics*, 2002, *33* (4), 689–708.

DeGroot, Morris H., "Reaching a Consensus," *Journal of the American Statistical Association*, 1974, *69* (345), 118–121.

Eggert, Wolfgang and Martin Kolmar, "Contests with size effects," *European Journal of Political Economy*, December 2006, *22* (4), 989–1008.

Esteban, Joan and Debraj Ray, "Conflict and Distribution," *Journal of Economic Theory*, August 1999, *87* (2), 379–415.

Esteban, Joan-María and Debraj Ray, "On the Measurement of Polarization," *Econometrica*, 1994, *62* (4), 819–851.

Funk, Peter, "Auctions with interdependent valuations," *International Journal of Game Theory*, March 1996, *25* (1), 51–64.

Galam, S., "Minority opinion spreading in random geometry," *The European Physical Journal B - Condensed Matter and Complex Systems*, February 2002, *25* (4), 403–406.

Golub, Benjamin and Matthew O Jackson, "Naïve Learning in Social Networks and the Wisdom of Crowds," *American Economic Journal: Microeconomics*, February 2010, *2* (1), 112–149.

Jehiel, Philippe, Benny Moldovanu, and Ennio Stacchetti, "How (Not) to Sell Nuclear Weapons," *The American Economic Review*, 1996, *86* (4), 814–829.

_ , _ , and _ , "Multidimensional Mechanism Design for Auctions with Externalities," *Journal of Economic Theory*, April 1999, *85* (2), 258–293.

Klose, Bettina and Dan Kovenock, "The all-pay auction with complete information and identity-dependent externalities," *Economic Theory*, 2015, *59* (1), 1–19.

Konrad, Kai A., "Silent interests and all-pay auctions," *International Journal of Industrial Organization*, July 2006, *24* (4), 701–713.

Lee, Sanghack, "Contests with size effects through costs," *European Journal of Political Economy*, December 2007, *23* (4), 1190–1193.

_ **and J. Hyeong Kang**, "Collective contests with externalities," *European Journal of Political Economy*, November 1998, *14* (4), 727–738.

Lehrer, Ehud and Dimitry Shaiderman, "Markovian Persuasion," November 2021. arXiv:2111.14365 [econ].

Linster, Bruce G., "A Generalized Model of Rent-Seeking Behavior," *Public Choice*, 1993, *77* (2), 421–435.

Moldovanu, Benny, Aner Sela, and Xianwen Shi, "Carrots and Sticks: Prizes and Punishments in Contests," *Economic Inquiry*, 2012, *50* (2), 453–462.

Mustafaraj, Eni, Samantha Finn, Carolyn Whitlock, and Panagiotis T. Metaxas, "Vocal Minority Versus Silent Majority: Discovering the Opionions of the Long Tail," in "2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing" October 2011, pp. 103–110.

Nitzan, Shmuel, "Rent-Seeking with Non-Identical Sharing Rules," *Public Choice*, 1991, *71* (1/2), 43–50.

Park, Sung-Hoon and Sanghack Lee, "Contests with Linear Externality in Prizes," August 2019.

Sela, Aner, "Optimal allocations of prizes and punishments in Tullock contests," *International Journal of Game Theory*, September 2020, *49* (3), 749–771.

Shaffer, Sherrill, "War, labor tournaments, and contest payoffs," *Economics Letters*, August 2006, *92* (2), 250–255.

Skaperdas, Stergios, "Contest Success Functions," *Economic Theory*, 1996, *7* (2), 283–290.

Tullock, Gordon, "Efficient rent seeking," in "Buchanan, J.M., Tollison, R.D., and Tullock, G., (Eds.) , Toward a Theory of the Rent-seeking Society" number 4. In '1.', College Station: Texas A&M University, 1980, pp. 97–112.

A Proof of Theorem 1

Proof. The zero vector is clearly not an equilibrium, so fix a non-zero profile $\mathbf{e} \in \mathbb{R}^n_+$, and consider the payoff function of player $j \in N_i$,

$$\begin{aligned} U_j(e_j, e_{-j}|O_i) &= -e_j - \frac{\sum_{l=1}^{k} \sum_{r \in N_l} e_r |O_l - O_i|}{\sum_{r=1}^{n} e_r} \\ &= -e_j - \frac{\sum_{l=1}^{k} E_l |O_l - O_i|}{\sum_{l=1}^{k} E_l}. \end{aligned}$$

The function $U_j(\cdot, e_{-j}|O_i)$ is differentiable and concave in e_j, so the maximum is reached either at the boundary $e_j = 0$ (effort levels are unbounded from above), or when the following FOC is satisfied:

$$\frac{\partial U_j(e_j, e_{-j}|O_i)}{\partial e_j} = \sum_{l=1}^{k} E_l |O_l - O_i| - \left[\sum_{l=1}^{k} E_l\right]^2 = 0, \quad \forall j = 1, \ldots, n.$$

Denote $d_{l,i} = |O_l - O_i|$, and note that

$$d_{l,i} - d_{l,i+1} = |O_l - O_i| - |O_l - O_{i+1}| = \begin{cases} -d_{i,i+1}, & \forall l \leq i, \\ d_{i,i+1}, & \forall l > i. \end{cases}$$

For every $i = 1, \ldots, k-1$, compute the difference

$$\begin{aligned} \frac{\partial U_j(\mathbf{e}|O_i)}{\partial e_j} - \frac{\partial U_{j'}(e|O_{i+1})}{\partial e_{j'}} &= \sum_{l=1}^{k} E_l d_{l,i} - \sum_{l=1}^{k} E_l d_{l,i+1} \\ &= -\sum_{l \leq i} E_l d_{i,i+1} + \sum_{l > i} E_l d_{i,i+1} = 0. \end{aligned} \tag{2}$$

Divide every such Equation 2 (for opinion O_i) by $d_{i,i+1} \neq 0$ to get

$$H_i := -\sum_{l \leq i} E_l + \sum_{l \geq i+1} E_l = 0.$$

Subtract $H_{i-1} - H_i$ to get $2E_i = 0$ for every $i = 2, \ldots, k-1$. Since effort levels are non-negative, we deduce that, in equilibrium, the first-order conditions are satisfied at the boundary $e_j = 0$, for every moderate player j. Thus, we are left with the following FOCs for the extreme opinions

$$E_i |O_1 - O_k| - [E_1 + E_k]^2 = 0, \text{ where } i = 1, k.$$

Solving for E_1 and E_k, we get a unique solution (other than the zero-effort profile) of $E_1 = E_k = \frac{|O_1 - O_k|}{4}$, as needed. □

B Proof of Theorem 2

Proof. Consider an equilibrium profile $\mathbf{e} \in \mathbb{R}^n_+$. It follows from the proof of Lemma 1 that $E_1 = E_3$, so the polarization level translates to

$$P(\mathbf{e}) \;=\; \frac{2E_1E_2^2 + 2E_1^2E_2 + 4E_1^3}{[2E_1 + E_2]^3} = \frac{W^2 + \frac{1}{2}W + \frac{1}{2}}{[1+W]^3} = \frac{1}{1+W} - \frac{1}{2(1+W)^2} - \frac{W}{(1+W)^3},$$

where $W = \frac{E_2}{2E_1}$. According to Lemma 1, W is strictly decreasing in δ, so it is left to prove that $P(\mathbf{e})$ is strictly decreasing w.r.t. $W \geq 0$. Evidently,

$$\frac{dP}{dW} = -\frac{1}{(1+W)^2} + \frac{3W}{(1+W)^4} = \frac{-W^2 + W - 1}{(1+W)^4} < 0,$$

for every $W \geq 0$, as needed. □

C Proof of Lemma 1

Proof. Fix $\delta \in (0,1)$ and consider the FOCs of every player $j \in N_i$ given a non-zero profile e,

$$\sum_{l=1}^{3} E_l \delta^{|i-l|} |O_l - O_i| = \left[\sum_{l=1}^{3} \delta^{|i-l|} E_l\right]^2,$$

where $E_l = \sum_{r \in N_l} e_r$ for $1 \leq l \leq 3$. Stated explicitly for every opinion, we get

$$\begin{aligned}
\text{for } j \in N_1 \quad &: \quad E_2\delta + 2E_3\delta^2 = \left[E_1 + E_2\delta + E_3\delta^2\right]^2, \\
\text{for } j \in N_2 \quad &: \quad E_1\delta + E_3\delta = \left[E_1\delta + E_2 + E_3\delta\right]^2, \\
\text{for } j \in N_3 \quad &: \quad 2E_1\delta^2 + E_2\delta = \left[E_1\delta^2 + E_2\delta + E_3\right]^2.
\end{aligned}$$

Define $X = E_1 + E_2\delta + E_3\delta^2$, $Y = E_1\delta + E_2 + E_3\delta$, and $Z = E_1\delta^2 + E_2\delta + E_3$. So,

$$\begin{aligned}
X - E_1 + \delta^2 E_3 &= X^2, \\
Y - E_2 &= Y^2, \\
Z - E_3 + \delta^2 E_1 &= Z^2.
\end{aligned}$$

and

$$\begin{aligned}
X - \delta Y = E_1(1-\delta^2) \quad &\Rightarrow \quad E_1 = \frac{X - \delta Y}{1-\delta^2}, \\
Z - \delta Y = E_3(1-\delta^2) \quad &\Rightarrow \quad E_3 = \frac{Z - \delta Y}{1-\delta^2}.
\end{aligned}$$

Plug this in the previous equations to get

$$X^2 = X - \frac{X-\delta Y}{1-\delta^2} + \delta^2\frac{Z-\delta Y}{1-\delta^2} = X + \frac{\delta^2 Z - X}{1-\delta^2} + \delta Y \quad \Rightarrow \quad X^2(1-\delta^2) = (Z-X)\delta^2 + \delta(1-\delta^2)Y,$$
$$Z^2 = Z - \frac{Z-\delta Y}{1-\delta^2} + \delta^2\frac{X-\delta Y}{1-\delta^2} = Z + \frac{\delta^2 X - Z}{1-\delta^2} + \delta Y \quad \Rightarrow \quad Z^2(1-\delta^2) = (X-Z)\delta^2 + \delta(1-\delta^2)Y.$$

Subtracting both equations yields $(X^2 - Z^2)(1-\delta^2) + 2(X-Z)\delta^2 = 0$. Hence, we conclude that $X = Z$ is the unique solution and $E_1 = E_3$.

So, the FOCs revert to

$$2\delta^2 + 2\delta W = E_1\left[1 + \delta^2 + 2\delta W\right]^2,$$
$$2\delta = E_1\left[2\delta + 2W\right]^2,$$

where $W = E_2/(2E_1)$. Divide the first equation by the second to get

$$\delta + W = \left[\frac{1+\delta^2+2\delta W}{2(\delta+W)}\right]^2 \quad \Leftrightarrow \quad 4(\delta+W)^3 = \left[1+\delta^2+2\delta W\right]^2.$$

Define the function $Q(W,\delta) = 4(\delta+W)^3 - \left[1+\delta^2+2\delta W\right]^2$ and note that $Q(0,\delta) < 0$ and $Q(1,\delta) > 0$, for every $\delta \in (0,1]$. By the Intermediate Value Theorem, there exists a solution for $Q(W(\delta),\delta) = 0$. Note that

$$\begin{aligned}\frac{\partial Q}{\partial W} &= 12(\delta+W)^2 - 4W[1+\delta^2+2\delta W] \\ &\geq 12(\delta+W)^2 - 4(W+\delta)[1+\delta^2+2\delta W] = \frac{\partial Q}{\partial \delta},\end{aligned}$$

and by substituting $\left[1+\delta^2+2\delta W\right] = 2(\delta+W)^{3/2}$ we get

$$\begin{aligned}\frac{\partial Q}{\partial \delta} &= 12(\delta+W)^2 - 4[1+\delta^2+2\delta W](\delta+W) \\ &= 12(\delta+W)^2 - 8(\delta+W)^{5/2} \\ &= 8(\delta+W)^2(1.5 - \sqrt{\delta+W}) > 0, \quad \forall(\delta,W) \in (0,1]^2.\end{aligned}$$

Hence, both partial derivatives are strictly positive, and the solution $W(\delta)$ to $Q(W,\delta) = 0$ is unique. By the Implicit Function Theorem, we get

$$\frac{\partial W(\delta)}{\partial \delta} = -\frac{\frac{\partial Q}{\partial \delta}}{\frac{\partial Q}{\partial W}} < 0,$$

implying that $W = \frac{E_2}{E_1+E_3}$ is decreasing w.r.t. δ in equilibrium. □

D Proof of Lemma 2

Proof. Note that $\pi_1 + \pi_2 + \pi_3 = 2\pi_1 + \pi_2 = 1$, so it is sufficient to prove that $\frac{\pi_1}{\pi_2}$ is increasing in δ. Denote $D = \frac{\pi_1}{\pi_2} = \frac{r^*}{\sqrt{\delta + \frac{1}{2}r^*}}$ and differentiate with respect to δ, so that

$$\frac{\partial D}{\partial \delta} = \frac{r^* - \frac{dr^*}{d\delta}[2\delta + \frac{1}{2}r^*]}{2(r^*)^2\sqrt{\delta + \frac{1}{2}r^*}}.$$

Since r^* is decreasing in δ, we get $\frac{\partial D}{\partial \delta} > 0$, and the result holds. □