Homo moralis goes to the voting booth: coordination and information aggregation

Ingela Alger, Jean-François Laslier

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

WORKING PAPER N° 2020 – 78

Homo moralis goes to the voting booth: coordination and information aggregation

Ingela Alger
Jean-François Laslier

JEL Codes:
Keywords: voting, Homo moralis, Kantian morality, social dilemmas, divided majority problem, Condorcet jury theorem

Funded by a French government subsidy managed by the ANR under the framework of the Investissements d'avenir programme reference ANR-17-EURE-001

Homo moralis goes to the voting booth: coordination and information aggregation*

Ingela Alger† Jean-François Laslier‡

November 30, 2020

Abstract

This paper revisits two classical problems in the theory of voting—viz. the divided majority problem and the strategic revelation of information by majority vote—in the light of evolutionarily founded partial Kantian morality. It is shown that, compared to electorates consisting of purely self-interested voters, such Kantian morality helps voters solve coordination problems and improves the information aggregation properties of equilibria, even for modest levels of morality.

Keywords: voting, *Homo moralis*, Kantian morality, social dilemmas, divided majority problem, Condorcet jury theorem

1 Introduction

The question of individual cooperation is a puzzle for social theories because cooperation should be sustained when efficient for the group but might be in contradiction with efficiency at the individual level. This puzzle appears under various disguises in different disciplines:

*I.A. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC EvolvingEconomics), as well as IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program). J.-F.L. thanks the support of the EUR grant ANR-17-EURE-0001. We are grateful for extensive discussions with Jörgen Weibull, as well as comments from audiences at the Seoul Social Choice and Welfare Society meeting, the Paris School of Economics, and the COMSOC on line seminar. Olivier Lision provided excellent research assistance.

†Toulouse School of Economics, CNRS, University of Toulouse Capitole, and Institute for Advanced Study in Toulouse, France. ingela.alger@tse-fr.eu

‡Paris School of Economics (CNRS). jean-francois.laslier@ens.fr

Evolutionary Biology (Nowak and Sigmund 2005 [29]), Ethology (de Waal 1996 [39]), Economics (Moulin 1995 [25]), Political theory (Ostrom 1998 [30]) or Social Philosophy (Binmore 1994 [8]). In the light of recent results from the literature on the evolutionary foundations of human motivation, we use formal game theory to revisit two classic cooperation dilemmas faced by voters whose ability to communicate with each other is limited: the divided majority problem and the Condorcet jury theorem.

Grounded on evolutionary considerations, Alger and Weibull (2013 [2]) have proposed the mathematical model of morality, termed *Homo moralis*, that will be used in this paper. *Homo moralis* reasons as follows: when contemplating a course of action, she evaluates what her material payoff would be if each other individual of the population she belongs to, were to follow the same course of action with probability κ. Although one might dispute whether this behavior captures the whole significance of Immanuel Kant's morality, it incorporates a key ingredient of this construct (and, arguably, of most moral theories): the "universalization" principle (Kant 1785 [19], Roemer 2019 [32]). For $\kappa = 0$, the individual is purely concerned by her material payoff, but for $\kappa = 1$, the individual decides to undertake an action if this action has good consequences *once adopted by everyone*. In this sense, *Homo moralis* is a model of partial Kantian morality, and the parameter κ is interpreted as a level of morality. So the model also stands on its own feet, independently of its evolutionary foundations. It can be read as modeling a particular "ethical" behavior based on partial universalization.

Since Jean-Jacques Rousseau (1755 [33]), many scholars have expressed the idea that political psychology should not be cut from its possible biological roots: see for instance Shubert (1982 [34]), Petersen (2015 [31]), Sidanus and Kurzban (2013 [36]) or Bergner and Hatemi (2017 [7]). Within this stream of research, *Homo moralis* is a theoretical model that does not link to empirical genetics but to the pure theory of evolution and stability of interactive behavior. As in Economics (Lesourne et al. 2006 [23]) the evolutionary approach complements the now-standard but still criticized theory of rational choice (Downs 1957 [11], Green and Shapiro, 1994 [18], Cox 1997 [10], Stephenson et al. 2018 [37]), by refining the concept of Nash equilibrium (which does not contain by itself notions of stability or convergence).

The evolutionary argument that implies precisely the behavior termed "*Homo moralis*" rests on the ideas that at least part of the fitness an individual achieves depends on the material payoff she achieves in social interactions, that her subjective utility (whose maximization drives individual behavior) is transmitted to her (biological or cultural) offspring, and that when a mutant utility function appears in the population its carriers are more exposed to interaction with other mutants than are non-mutants (because interactions are local). From

these premises, Alger and Weibull 2013 [2] showed that the toolkit of evolutionary game theory (Maynard Smith 1982 [24]) can be used to prove that evolutionarily stable preferences are of the *Homo moralis* type.

In this paper we take this result as our starting point and study the consequences of evolutionary Kantian morality in two distinct settings: first when voters face a pure coordination problem, and second when voters face an information aggregation problem.[1] Interest in these questions is warranted for many reasons, in particular because answers to these questions are necessary to properly evaluate voting rules from a normative point of view. But the theory of voting, on top of being of political relevance *per se*, contains the study of several archetypal situations of interaction, or "games" that are of broad interest. For instance the political game of coordination of a divided majority can be seen as a toy model of social coordination in general.[2]

The paper is organized as follows. In section 2 we formally describe the *Homo moralis* model. Then we consider two classical problems in the theory of voting: the divided majority problem under plurality rule in section 3 and the question of strategic revelation of information in the Condorcet jury setting in section 4. For each of these questions we study the implications of the hypothesis of evolutionary Kantian morality. The last section is a short conclusion.

2 Who is *Homo moralis*?

In this section we provide a formal definition of *Homo moralis* preferences. We start by the simple case of a two-player ($n = 2$) symmetric normal form game. Let X denote the set of pure strategies and $\pi(x, y)$ the material payoff for a player playing x in his interaction with a player playing y. Then a *Homo moralis* with degree of morality $\kappa \in [0, 1]$ achieves the following *utility* from using strategy x when the opponent uses strategy y:

$$U(x, y) = (1 - \kappa) \cdot \pi(x, y) + \kappa \cdot \pi(x, x). \tag{1}$$

[1] The question of participation when the electorate is large and voting is costly is tackled in a companion paper [1].

[2] Interestingly, the literature on animal behavior describes several collective phenomena similar to voting, from Honeybees to African wild dogs, relying on the interpretation of various techniques of social communication. See Walker et al. 2017 [40], Seeley 2010 [35], Sumpter 2010 [38].

The first term is the individual's material payoff, given the strategies effectively used. The second term captures the Kantian moral concern: it induces the individual to ponder what his/her material payoff would be if, hypothetically, the other individual were to use the same strategy as him/her. A *Homo moralis* with degree of morality κ thus chooses a strategy that maximizes the weighted sum of own material payoff and the Kantian moral concern, the weight attached to the latter being κ. In his *Grundlegung zür Metaphysik der Sitten* (1785), Immanuel Kant wrote "Act only according to that maxim whereby you can at the same time will that it should become a universal law." In this vein, *Homo moralis* can be said to "act according to that maxim whereby you can at the same time will that others should do likewise with some probability."

The probability interpretation is particularly compatible with the logic whereby *Homo moralis* preferences have been shown to have a strong evolutionary foundation. The argument is as follows. Consider a large population where each individual inherits his/her preferences from an individual in the preceding generation (be it culturally or biologically), and in each generation individuals are matched at random to interact in pairs according to the material game described above. *Homo moralis* preferences are justified based on the observation that in essentially all populations new cultural variants or genetic mutations spread locally. From this fact it follows that, even if the probability of two similar mutants being matched is very small, a mutant is still relatively more likely than non-mutants to be matched with a mutant. Taking into consideration this phenomenon, Alger and Weibull ([2]) show that, for a value of κ that precisely equals the probability that mutants are matched when mutants are vanishingly rare, individuals who maximize utility of the *Homo moralis* form have an evolutionary advantage over those who would behave differently, when preferences are passed on from one generation to the next and the number of individuals to whom an individual passes his/her preferences is determined by the material payoff (s)he obtains.

We turn now to the more general case of an n-player interaction ($n \geq 2$). The material payoff of a player i depends on her own strategy $x_i \in X$ and on the strategies $y_1, \ldots, y_{n-1}$ used by the other players. Writing $(y_1, \ldots, y_{n-1}) \equiv \boldsymbol{y}_{-i}$, the material payoff is denoted $\pi(x_i, \boldsymbol{y}_{-i})$. Let the symbol $\mathbb{E}$ denote mathematical expectation.

Definition 1 *In a symmetric n-player game π, an individual is a* ***Homo moralis*** *with degree of morality κ if his or her utility function U satisfies $U(x, \boldsymbol{y}) \equiv \mathbb{E}[\pi(x, \tilde{\boldsymbol{y}})]$ where $\tilde{\boldsymbol{y}} = (\tilde{y}_1, ..., \tilde{y}_{n-1})$ is a random strategy profile for the other players, with each component $\tilde{y}_i$ being the actual strategy used by opponent i (y_i) with probability $1 - \kappa$ and the individual's own strategy (x) with probability κ.*

For instance, in an interaction between three individuals the utility of a *Homo moralis* with degree of morality $\kappa \in [0,1]$ from using strategy x when the others use strategies y_1 and y_2 is:

$$U(x, y_1, y_2) = (1-\kappa)^2 \cdot \pi(x, y_1, y_2) + \kappa(1-\kappa) \cdot [\pi(x, x, y_2) + \pi(x, y_1, x)] + \kappa^2 \cdot \pi(x, x, x).$$

The two voting problems which will be studied in this article are cases of many-player interactions (the population of voters) who have access to the same strategies (the possible ballots to cast). A key feature of voting games is that they are *aggregative*, in the sense that (a) any individual's payoff depends only on how he/she votes and the vector of voting strategies played by the other individuals, and (b) the individual's payoff would not be affected if other individuals swapped their strategies. For instance the outcome of the vote only depends on the total number of votes obtained by each candidate. In the study of equilibrium, it will be sufficient for our purposes to state the utility that a *Homo moralis* with degree of morality κ achieves from playing strategy x when all the others use the same strategy, say y. In an aggregative game, this simplifies the writing of the general κ-moral utility function specified in Definition 1 to the following expression:

$$U\left(x, \mathbf{y}^{(n-1)}\right) = \sum_{m=1}^{n} \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \kappa^{m-1} (1-\kappa)^{n-m} \pi\left(x, \mathbf{x}^{(m-1)}, \mathbf{y}^{(n-m)}\right), \tag{2}$$

where $\mathbf{y}^{(\ell)}$ is the ℓ-dimensional vector whose components all equal y and $\mathbf{x}^{(\ell)}$ the ℓ-dimensional vector whose components all equal x. When all the other individuals use strategy y, the random strategy profile $\tilde{\boldsymbol{y}} = (\tilde{y}_1, ..., \tilde{y}_{n-1})$ in the general definition is such that each component other than the first one (which is the individual's own strategy) is a random variable that follows a binomial distribution, taking the value y with probability $1-\kappa$ and the value x with probability κ. Since the game is aggregative, one needs only keep track of the number of times that exactly m out of the $n-1$ components in $\tilde{\boldsymbol{y}}$ take the value x.

In one of the models below we will consider infinitely large populations, modeled as a continuum. To study equilibria in this setting we will take the utility of *Homo moralis* who plays strategy x in a population where all others play strategy y to be the material payoff should a share κ of the population (hypothetically) play x instead of y. We rely on the de Moivre-Laplace theorem to argue that this is a good approximation of the expression in (2) when n tends to infinity. Indeed, this theorem says that the probability mass function of the random number of times that y is replaced by x in n independent trials converges to the probability density function of the normal distribution with mean $n\kappa$ and standard deviation

$\sqrt{n\kappa(1-\kappa)}$ as $n \to \infty$.

Throughout the paper we use the Nash equilibrum concept: the originality of the *Homo moralis* model is to introduce a distinction between material payoff and utility, but it is amenable to application of the standard Nash notion. A Nash equilibrium strategy profile is a vector of strategies such that each player uses a strategy that maximizes her utility, given the strategies used by the others. We will say that a Nash equilibrium is *strict* if each player would obtain a strictly lower utility by any deviation from her strategy, *partially strict* if at least some players would obtain a strictly lower utility by some deviation from their strategy, and *flat* if all players are indifferent between deviating or not.

3 Does *Homo moralis* vote strategically in the divided majority model?

We first tackle the question of strategic voting in the divided majority setting. This is a coordination game between players that form a majority but lose the election if they split their votes among two candidates of their camp.

3.1 The divided majority model

Consider an infinite population (a continuum of mass one) of voters who are to elect one candidate. There are two political parties, A and B, and three candidates: one from party B, whom we simply call B and two from party A, whom we call $A1$ and $A2$. Some electors of party A prefer candidate $A1$ over $A2$ while others prefer $A2$ over $A1$. Letting n_C denote the share of the population that prefers candidate $C \in \{A_1, A_2, B\}$, where $n_{A1} + n_{A2} + n_B = 1$, we adopt the following assumption:

$$0 < n_{A2} < n_{A1} < n_B < n_{A1} + n_{A2} < 1. \tag{3}$$

In other words, candidate B is supported by a minority of size $n_B < 1/2$, while a majority of size $n_{A1} + n_{A2} > 1/2$ would prefer a candidate from party A to win over candidate B. However, the majority is divided into two groups, each of them smaller than the minority. Note that assumption (3) implies

$$1/3 < n_B < 1/2 \tag{4}$$

and n_B can take any value within these bounds. The following table shows the (material) payoff that a voter gets depending on which candidate is elected and which candidate (s)he prefers, where $\varepsilon \in (0,1)$ is a parameter that measures the disagreement between $A1$- and $A2$-supporters.

	$A1$	$A2$	B
$A1$	$1+\varepsilon$	$1-\varepsilon$	0
$A2$	$1-\varepsilon$	$1+\varepsilon$	0
B	0	0	1

We examine whether voters may be expected to vote sincerely—i.e., for their preferred candidate—or strategically—i.e., for a candidate other than the one they prefer. We concentrate on supporters of party A, who face a coordination problem. This is without loss of generality: B-supporters clearly have no incentive to deviate from voting for B. Under plurality voting, if A-supporters vote sincerely, the minority wins (B is elected); however, the majority can win by coordinating their votes on either $A1$ or $A2$, a coordination that requires some voters to vote strategically. Myerson and co-authors [15, 26, 27] use this game to show that Approval Voting can help solve this dilemma between strategic and sincere voting, and Myerson and Weibull (2015 [28]) take a similar game as a case-study in their theory of coordination. Here we will show how Kantian morality in the form of *Homo moralis* preferences can help solve the dilemma under plurality voting.

Recall that *Homo moralis* can be said to "act according to that maxim whereby you can at the same time will that others should do likewise with some probability."Defining *Homo moralis* preferences precisely thus requires defining who the "others" are. We distinguish between two scenarios, the *ex post* and the *ex ante* one. In the *ex ante* scenario the voter does not know yet if his preferred candidate is $A1$ or $A2$, and the reference population is the whole population of A-voters; in the *ex post* scenario, the piece of information is known and the reference population is the group of $A1$-supporters for an $A1$-supporter, and the group of $A2$-supporters for an $A2$-supporter.

In all cases, in this section we consider a large population (a continuum of voters) so that the result of the election is deterministic, defined by the fractions of the population that vote for each candidate. Our goal being to characterize symmetric equilibria, we rely on the approximation of the utility in (2) described in Section 2 for infinitely large populations. Here the material payoff of a voter is $1+\varepsilon$, $1-\varepsilon$, or 0, depending on whether it is his preferred A-candidate, the other A-candidate, or candidate B who wins. Hence, for a given strategy played by all others in his reference population, a *Homo moralis* voter with degree of morality

κ evaluates each strategy by pondering what his material payoff would be if, hypothetically, a share κ of the voters in the reference population would also use this strategy.

3.2 The divided majority: the *ex post* scenario

In the *ex post* scenario, all the A-supporters first learn their ranking of candidates $A1$ and $A2$, and if they have *Homo moralis* preferences with some positive degree of morality κ they use the group of voters who have the same ranking over these candidates as the reference group. Thus, for a voter who prefers candidate $A1$ (resp. $A2$) the reference group is the n_{A1} (resp. n_{A2}) voters who also prefer $A1$ (resp. $A2$).

We first examine whether sincere voting is an equilibrium. By sincere voting we mean the situation in which all voters vote for their preferred candidate. Sincere voting leads to the election of B. For $\kappa = 0$, this is a flat equilibrium, because each voter only considers how her action impacts her material payoff, the material payoff only depends on who is elected, and, with a continuum of voters none of them is pivotal. Turning now to *Homo moralis* preferences with a positive degree of morality κ, this conclusion does not necessarily hold, as shown in the following proposition.[3]

Proposition 1 (Ex post sincere voting) *Suppose that all the voters are Homo moralis with degree of morality* $\kappa \in [0, 1]$. *Let* $\kappa^* = \frac{n_B - n_{A1}}{n_{A2}}$. *Then* $0 < \kappa^* < 1$ *and in the ex post scenario sincere voting (*$A1$*-supporters vote for* $A1$ *and* $A2$*-supporters for* $A2$*) is:*

- *a flat Nash equilibrium if* $\kappa < \kappa^*$*;*
- *not a Nash equilibrium if* $\kappa > \kappa^*$.

When the degree of morality is high enough, sincere voting is not sustainable because either some or all voters from the divided majority then prefer to vote strategically. Although such a deviation has no effect on the actual outcome, it brings satisfaction to a sufficiently moral *Homo moralis* to know that candidate B would be beaten, should a share κ of the voters in his reference group also vote strategically. Because $A1$-voters are more numerous than $A2$-voters, for any given degree of morality the deviation to strategic voting by a $A2$-voter is more effective than a deviation to strategic voting by a $A1$-voter, because B's advantage is

[3]For expositional simplicity, throughout we disregard any knife-edge case where κ exactly equals the threshold value at hand. Clearly, the set of equilibria for such knife-edge cases would depend on the assumption we would then have to make about tie-breaking rules, and this is not related to our argument.

the smallest when $A1$-voters vote sincerely. The threshold value κ^* is the degree of morality above which $A2$-voters strictly prefer to vote strategically vote for $A1$ rather than voting for their preferred candidate $A2$, given that $A1$-voters cast their ballots for candidate $A1$.

Strong enough Kantian moral concerns also enable coordination of the divided majority as a strict equilibrium, as shown next. Notice that, in this *ex post* scenario, strict coordination occurs either on $A1$ or on $A2$, with the same threshold value for κ.

Proposition 2 (Ex post coordination) *Suppose that all the voters are Homo moralis with degree of morality* $\kappa \in [0,1]$. *Let* $\kappa_1^{**} = \frac{n_A - n_B}{n_{A1}}$ *and* $\kappa_2^{**} = \frac{n_A - n_B}{n_{A2}}$. *Then* $0 < \kappa_1^{**} < \kappa_2^{**} < 1$ *and in the ex post scenario coordination on candidate* Ak, $k \in \{1,2\}$, *is:*

- *a flat Nash equilibrium if* $\kappa < \kappa_1^{**}$;
- *a partially strict Nash equilibrium if* $\kappa_1^{**} < \kappa < \kappa_2^{**}$;
- *a strict Nash equilibrium if* $\kappa > \kappa_2^{**}$.

In our model of the divided majority in a continuum population of voters, each individual voter has no real effect on the election outcome and derives no utility from expressing their opinion. Intuition thus suggests that each voter might as well vote on a random candidate. Remarkably, this is not true under *Homo moralis* preferences. As shown in the proposition, sufficiently pronounced Kantian moral concerns break the indifference and induce a strict preference to vote for one of the majority candidates, should all other majority supporters do the same. *Homo moralis* preferences thus allow the divided majority to sustain coordination on one of the A-candidates and win the election against B.

3.3 The divided majority: the *ex ante* scenario

Here we consider the *ex ante* situation, where each A-voter knows that she prefers party A, that there are three candidates $C = B, A1, A2$ with respective supports n_C, as well as the associated payoffs, but does not yet know if she will prefer $A1$ or $A2$. Unlike in the *ex post* setting, the reference population for an A-voter is the whole population of A-voters. This is not an artificial situation: it models, for instance, the reasoning of a citizen who wonders whether it is better to vote according to his ideas (voting sincerely, or naively for the candidate she will happen to prefer) or to (strategically) coordinate her vote with voters of the same camp.

For a supporter of party A, a (behavior) strategy specifies the candidate to vote for, depending on whether (s)he prefers candidate $A1$ or candidate $A2$. We will write such a strategy $\alpha = (\alpha_1, \alpha_2)$; for instance

$$\alpha^{\text{si}} = (A1, A2)$$

is the "sincere" strategy and

$$\alpha^{(1)} = (A1, A1)$$

is the strategy which dictates to vote $A1$ in any case. In what follows we will not consider the reversed strategy (α_1, α_2).[4] Hence, there are three strategies: $\alpha^{(1)}$, $\alpha^{(2)}$ and α^{si}.

Let $C^*(\alpha)$ denote the candidate that wins given that all A-voters use strategy α. Then, the following expression shows the expected material payoff of a A-voter when all other A-voters use α.

$$\pi_A(\alpha) = \begin{cases} 1 + \varepsilon \cdot (n_{A1} - n_{A2})/(n_{A1} + n_{A2}) & \text{if } \; C^*(\alpha) = A1 \\ 1 - \varepsilon \cdot (n_{A1} - n_{A2})/(n_{A1} + n_{A2}) & \text{if } \; C^*(\alpha) = A2 \\ 0 \;\; \text{if } \; C^*(\alpha) = B. \end{cases} \tag{5}$$

In words, from behind the veil of ignorance as to whether (s)he prefers candidate $A1$ or $A2$, a voter with ranking A gets material payoff that is sometimes slightly above 1 and sometimes slightly below 1 if an A-candidate wins. Since (s)he is more likely to prefer $A1$ to $A2$ (i.e., $n_{A1} > n_{A2}$) (s)he would prefer $A1$ to win from an *ex ante* perspective. Such a voter is, moreover, certain to get material payoff 0 if B wins. Note that since there is a continuum of voters the winning candidate does not depend on the strategy used by the individual voter at hand.

Using $C^{(\kappa)}(\alpha', \alpha)$ to denote the candidate that would win should a share κ of the other A-voters use the same strategy as him/her (α') instead of the strategy that they are actually using (α), the *utility* of a *Homo moralis* with degree of morality κ is:

$$U_A^{(\kappa)}(\alpha', \alpha) = \begin{cases} 1 + \varepsilon \cdot (n_{A1} - n_{A2})/(n_{A1} + n_{A2}) & \text{if } \; C^{(\kappa)}(\alpha', \alpha) = A1 \\ 1 - \varepsilon \cdot (n_{A1} - n_{A2})/(n_{A1} + n_{A2}) & \text{if } \; C^{(\kappa)}(\alpha', \alpha) = A2 \\ 0 \;\; \text{if } \; C^{(\kappa)}(\alpha', \alpha) = B. \end{cases} \tag{6}$$

We next turn to analysis of the following two questions: (a) under which conditions is sincere voting a Nash equilibrium? (b) under which conditions is strategic voting—whereby the divided majority achieves the election of either candidate $A1$ or candidate $A2$—a Nash

[4]The reader will easily check that adding this possibility does not alter the results.

equilibrium? As a benchmark, consider first briefly the (standard) case in which voters care only about their material payoffs (i.e., they have *Homo moralis* preferences with $\kappa = 0$). Since each individual vote has no impact on the outcome, any strategy profile—both sincere and strategic voting—is a Nash equilibrium. Indeed, given that all other voters play a certain strategy, any given voter is indifferent between all the available strategies. Turning next to *Homo moralis* preferences, we will see how these eliminate certain equilibria when the degree of morality is pronounced enough.

We first show that sincere voting is not necessarily a Nash equilibrium under *Homo moralis* preferences.

Proposition 3 (Ex ante sincere voting) *Suppose that all the voters are Homo moralis with degree of morality $\kappa \in [0, 1]$. For the same $\kappa^* = \frac{n_B - n_{A1}}{n_{A2}}$ as in the ex post scenario (see Proposition 1), in the ex ante scenario, sincere voting (the strategy profile $\alpha^{\text{si}} = (A1, A2)$) is:*

- *a flat Nash equilibrium if $\kappa < \kappa^*$;*
- *not a Nash equilibrium if $\kappa > \kappa^*$.*

As the reader can see, with respect to sincere voting the argument as well as the result in the *ex ante* case are strictly identical to the ones in the *ex post* case (see Proposition 1).

Secondly, we show that strategic voting whereby A-voters coordinate on candidate $A1$, is sustainable in a strict sense as a Nash equilibrium when A-voters have *Homo moralis* preferences with a sufficiently high degree of morality.

Proposition 4 (Ex ante coordination on $A1$) *Suppose that all the voters are Homo moralis with degree of morality $\kappa \in [0, 1]$. For the same $\kappa_2^{**} = \frac{n_A - n_B}{n_{A2}}$ as in Proposition 2, in the ex ante scenario coordination on candidate $A1$ (the strategy profile $\alpha^1 = (A1, A1)$) is:*

- *a flat Nash equilibrium if $\kappa < \kappa_2^{**}$;*
- *a strict Nash equilibrium if $\kappa > \kappa_2^{**}$.*

Comparing this proposition to Proposition 2, we see that coordination on $A1$, the strongest A-candidate, obtains for the same degrees of morality in the two scenarios. Such is not the case for coordination on $A2$, the weak A-candidate, as shown next.

Proposition 5 (Ex ante coordination on $A2$) *Suppose that all the voters are Homo moralis with degree of morality $\kappa \in [0,1]$. Let $\kappa^{***} = \frac{n_A - n_B}{n_A}$ and $\kappa^{****} = \frac{n_B}{n_A}$. Then $0 < \kappa^{***} < 1/2 < \kappa^{****} < 1$ and, in the ex ante scenario, coordination on the candidate $A2$ (the strategy profile $\alpha^2 = (A2, A2)$) is:*

- *a flat Nash equilibrium if $\kappa < \kappa^{***}$;*
- *a strict Nash equilibrium if $\kappa^{***} < \kappa < \kappa^{****}$;*
- *not a Nash equilibrium if $\kappa > \kappa^{****}$.*

It is more challenging to obtain coordination on $A2$ than on $A1$, because an A-supporter is more likely to end up being an $A1$- than an $A2$-supporter. Hence, an *ex ante* commitment to vote for $A2$ entails a sacrifice in terms of expected utility, which is not sustainable for degrees of morality so large that a deviation to a commitment to vote for $A1$ instead entails a hypothetical victory for candidate $A1$.

3.4 Conclusion on the divided majority problem

Summing up the insights generated by the analysis above, for low degrees of morality κ, any strategy profile is a non-strict equilibrium, following the "ocean of voters" logic: what I do personally does not matter and the same remains true if I have only a small number of (hypothetical) followers. Hence, for low degrees of morality no theoretical prediction arises. By contrast, for κ large enough, Kantian morality solves the coordination problem in the divided majority setting: sincere voting (which means non-coordination) is no longer an equilibrium, but coordination becomes a strict equilibrium. In the *ex post* scenario, coordination is sustained in the same way on any one of the two A-candidates, while in the *ex ante* scenario, coordination on the strongest of the two A-candidates is more readily obtained. This is summarized in the following theorem, that immediately follows from the preceding propositions.

Theorem 1 *In the divided majority problem, suppose that all the voters are Homo moralis with degree of morality $\kappa \in [0,1]$. Let $\kappa^* = \frac{n_B - n_{A1}}{n_{A2}}$, $\kappa^{**} = \frac{n_A - n_B}{n_{A2}}$, $\kappa^{***} = \frac{n_B}{n_A}$.*

- *Sincere voting, in both the ex post and ex ante scenarios, is a flat Nash equilibrium if $\kappa < \kappa^*$ and is not a Nash equilibrium if $\kappa > \kappa^*$.*
- *Coordination on $A1$, the strongest A-candidate, in both the ex post and ex ante scenarios, is a Nash equilibrium for any κ and a strict one if $\kappa > \kappa^{**}$.*

- *Coordination on A2, the weakest A-candidate,*
  - *in the ex post scenario, is a Nash equilibrium for any κ and a strict one if $\kappa > \kappa^{**}$;*
  - *in the ex ante scenario, is a Nash equilibrium if $\kappa < \kappa^{***}$ and is not a Nash equilibrium if $\kappa > \kappa^{***}$.*

Having thus examined the coordination problem, we turn to the information aggregation problem.

4 Should Homo Moralis sit in the jury ?

4.1 The jury model

Consider a group or jury of $n = 2m + 1$ members. These persons have to take a binary decision, say 0 or 1, according to a simple majority rule. There are two states of Nature, also labeled 0 and 1. All jurors agree that the right decision in each state $\omega \in \{0, 1\}$ equals ω, but they do not know the state of Nature. For the sake of simplicity we suppose that the material payoff of a juror is 1 if the decision is correct and 0 if not. Each juror's expected material payoff is thus the probability of a correct decision.[5]

The jurors share the common prior belief that the state is ω with probability $\mu_\omega \in (0, 1)$, where $\mu_0 + \mu_1 = 1$. Each jury member i also receives a private "signal" $s_i \in \{0, 1\}$, a random variable that is positively correlated with ω:

$$\begin{cases} \Pr[s_i = 0 \mid \omega = 0] = p_0 = 1 - q_0 > 1/2 \\ \Pr[s_i = 1 \mid \omega = 1] = p_1 = 1 - q_1 > 1/2. \end{cases}$$

A player's pure strategy specifies her vote as a function of her signal s_i. The set of strategies X thus consists of the following four strategies:

$$\xi^{\text{inf}} : \begin{cases} 0 \mapsto 0 \\ 1 \mapsto 1 \end{cases} ; \quad \xi^{0} : \begin{cases} 0 \mapsto 0 \\ 1 \mapsto 0 \end{cases} ; \quad \xi^{1} : \begin{cases} 0 \mapsto 1 \\ 1 \mapsto 1 \end{cases} ; \quad \xi^{\text{inv}} : \begin{cases} 0 \mapsto 1 \\ 1 \mapsto 0 \end{cases}$$

The classical Condorcet jury theorem (Condorcet 1785 [9]) concludes that the majority decision is informatively efficient in a large jury. This conclusion is reached under the assumption

[5]Notice incidentally that this assumption has a straightforward interpretation in term of fitness as survival probability: If the group decision is correct, all members will survive with probability one, if not they all die. Think of honeybees that "vote" to choose where to locate their new hive.

that all jurors truthfully report their signals, that is they all use the informative strategy ξ^{inf}. However, as is known since Austen-Smith and Banks (1996 [5]), such a strategy profile is not necessarily a (Bayesian) Nash equilibrium. As a juror, if I believe that all the other jurors vote informatively, correct Bayesian reasoning makes me condition my vote on the event that the other jurors' vote are in a tie and makes my vote decisive. This is true even in a large jury, where the condition for the (improbable) event that the other votes are exactly in a tie becomes more informative than my own signal, which I should therefore neglect. This surprising result casts doubt on the efficiency of majority (and super-majority) rules as a procedure to aggregate information. It initiated an important literature, dealing with political elections, criminal juries, or board decisions (Feddersen and Pesendorfer, 1997, 1998 [13, 14], Gerardi and Yariv 2008 [16], Gersbach and Hann 2008 [17]).

Because the voter is conditioning on an event of low probability, the non-equilibrium conclusion is not robust to small variations of the model (Laslier and Weibull 2013 [22]). Moreover, careful analysis shows that the collective inefficiency resulting from individual rational Bayesian behavior is not that strong (Koriyama and Szentes 2009 [20]). Still, these game-theoretical analyses fall short of a justification of informative voting behavior in the jury setting. In this section we compare a jury whose jurors have *Homo moralis* preferences with a jury whose jurors have *Homo oeconomicus* preferences. We ask whether informative voting is a Bayesian Nash equilibrium for a larger set of parameter constellations in the former than in the latter jury. We will further examine whether this set coincides with that in which informative voting by all jury members is efficient in the sense that it maximizes the probability that a correct decision is made.

4.2 A jury with three members

For a single decision-maker (a "jury of $n = 1$ member") who holds a very dissymmetrical prior and/or receives signals of very low quality, it may well be rational to always vote for the most probable state ω, without taking into account the received signal. By Bayes' law:

$$Pr[\omega = 0|s_i = 0] = \frac{p_0\mu_0}{p_0\mu_0 + (1-p_1)\mu_1} \text{ and } Pr[\omega = 1|s_i = 0] = \frac{(1-p_1)\mu_1}{p_0\mu_0 + (1-p_1)\mu_1},$$

so that deciding for $\omega = 0$ upon receiving signal 0 is optimal if and only if $p_0\mu_0$ is larger than$(1-p_1)\mu_1$. When this condition and the symmetric one for state 1 are met, informative voting is efficient for the single decision-maker. This happens for moderate values of the

prior odds ratio:

$$\frac{1-p_1}{p_0} < \frac{\mu_0}{\mu_1} < \frac{p_1}{1-p_0}. \tag{7}$$

Likewise, in a jury with three members, the symmetric strategy profile according to which all jury members vote informatively, $\boldsymbol{\xi}^{\mathbf{inf}} \equiv (\xi^{\inf}, \xi^{\inf}, \xi^{\inf})$, yields a higher probability of a correct decision than the symmetric strategy profiles according to which all jury members vote 0 or all vote 1, if and only if

$$\frac{(1-p_1)^2}{p_0^2} \cdot \frac{1+2p_1}{3-2p_0} < \frac{\mu_0}{\mu_1} < \frac{p_1^2}{(1-p_0)^2} \cdot \frac{3-2p_1}{1+2p_0}. \tag{8}$$

These inequalities are obtained by comparing the probability of a correct decision under the three alternative symmetric strategy profiles, which is equal to μ_0 if all jurors play ξ^0, to μ_1 if all jurors play ξ^1, and

$$\mu_0[p_0^3 + 3p_0^2(1-p_0)] + \mu_1[p_1^3 + 3p_1^2(1-p_1)]$$

if all jurors play $\xi^{\inf}$.

Having dealt with the normative properties of informative voting, we turn to the following positive question: under what condition is $\boldsymbol{\xi}^{\mathbf{inf}}$ a Bayesian Nash equilibrium in a jury with three jurors? For comparison, and to prepare the ground for analysis of a jury composed of individuals with *Homo moralis*, we examine this question under the assumption that all the jurors are *Homo oeconomicus*, i.e., their utility coincides with their material payoff.

As is known since Austen-Smith and Banks (1996 [5]), to check whether $\xi^{\inf}$ is a best response for a *Homo oeconomicus* individual i to $(\xi^{\inf}, \xi^{\inf})$, it is necessary and sufficient to examine how a deviation to another strategy would affect i's material utility in states of the world where i is pivotal, since these are the only states where the deviation would affect the outcome of the vote. Since the other two jury members—call them j and k—play $\xi^{\inf}$, it is thus sufficient to compare the expected costs and expected benefits of deviating in states where j and k received different signals, $s_j \neq s_k$. In this subsection we will without loss of generality assume that $\omega = 0$ is the least likely state, i.e., that $\mu_0 < 1/2$.

Consider first a deviation by i from $\xi^{\inf}$ to ξ^1. Such a deviation alters the vote outcome from 0 to 1 if $s_j \neq s_k$ and $s_i = 0$. The change in the vote outcome raises the material utility if the state of Nature is $\omega = 1$ but lowers it if the state of Nature is $\omega = 0$. Hence, i strictly prefers not to deviate to ξ^1 if and only if the probability that $s_j \neq s_k$, $s_i = 0$, and $\omega = 0$,

exceeds the probability that $s_j \neq s_k$, $s_i = 0$, and $\omega = 1$:[6]

$$\mu_0 \cdot p_0 \cdot 2p_0(1-p_0) > \mu_1 \cdot (1-p_1) \cdot 2p_1(1-p_1).$$

Likewise, a deviation by i from ξ^{inf} to ξ^0 alters the outcome from 1 to 0 if $s_j \neq s_k$ and $s_i = 1$, and this raises the material utility if the state of Nature is $\omega = 0$ but lowers it if the state of Nature is $\omega = 1$. Hence, the condition for the deviation to ξ^0 to be unviable boils down to:

$$\mu_1 \cdot p_1 \cdot 2p_1(1-p_1) > \mu_0 \cdot (1-p_0) \cdot 2p_0(1-p_0).$$

In sum, $\boldsymbol{\xi^{\text{inf}}}$ is a strict Bayesian Nash equilibrium in a jury composed of three *Homo oeconomicus*, if and only if

$$\frac{(1-p_1)^2}{p_0^2} \cdot \frac{p_1}{1-p_0} < \frac{\mu_0}{\mu_1} < \frac{p_1^2}{(1-p_0)^2} \cdot \frac{1-p_1}{p_0}. \tag{9}$$

In view of this equation, we see that the individual rationality of informative voting in a group of three is neither implied by rationality of informative voting for a single individual (compare with equation (7)) nor by its efficiency for the group (compare with equation (8)).

We turn now to a jury composed of jurors with *Homo moralis* preferences with some degree of morality $\kappa \in [0, 1]$. The reasoning is similar to the one above, in the sense that it is necessary and sufficient to consider situations in which the deviation affects the outcome of the vote. Thus, like above, a juror who ponders a certain deviation evaluates his/her expected material payoff if the deviation affects the outcome of the vote because (s)he is pivotal. However, in addition, a *Homo moralis* also evaluates how the deviating strategy would affect his/her expected material payoff if, hypothetically, with probability κ each other juror were also to play this strategy instead of the one (s)he is actually playing. We will say that jury member i is κ-pivotal if a deviation by i from strategy ξ to some strategy ξ' would affect the outcome of the vote in the hypothetical scenarios envisaged by *Homo moralis* in which at least one of the other jury members also play ξ'.

Like above, consider again juror i, and assume that j and k play ξ^{inf}. Table 1 lists the signal combinations (s_i, s_j, s_k) for which i is either pivotal or κ-pivotal. The first three columns show the signals received by the three jury members. The fourth column shows the vote outcome if i plays ξ^{inf}. The last four columns display the outcome of the vote if i were to deviate to ξ^1, and each of these columns corresponds to a different scenario that

[6]Like Austen-Smith and Banks (1996 [5]) we disregard parameter constellations where jurors are indifferent between strategies and thus focus on strict inequalities.

Homo moralis envisages. Thus, the column labeled (a) is the outcome when i deviates to ξ^1 while j and k play $\xi^{\inf}$. In columns (b) (resp. (c)), i ponders what the outcome would be if, hypothetically, j but not k (resp. k but not j) played ξ^1 instead of $\xi^{\inf}$, a scenario to which a *Homo moralis* with degree of morality κ attaches weight $\kappa(1-\kappa)$. Finally, in the last column i ponders what the outcome would be if, hypothetically, both of the other jurors played ξ^1 instead of $\xi^{\inf}$, a scenario to which a *Homo moralis* with degree of morality κ attaches weight κ^2. The signal realizations (s_i, s_j, s_k) not listed in this table are irrelevant, because for all of them the outcome of the vote is 1, whether or not i deviates to ξ^1.

Table 1: Signal realizations (s_i, s_j, s_k) for which a deviation by i from $\xi^{\inf}$ to ξ^1 would alter the vote outcome, if: (a) j and k play $\xi^{\inf}$, (b) hypothetically, j were also to play ξ^1 instead of $\xi^{\inf}$, (c) hypothetically, k were also to play ξ^1 instead of $\xi^{\inf}$, (d) hypothetically, both j and k were also to play ξ^1 instead of $\xi^{\inf}$

				(a)	(b)	(c)	(d)
s_i	s_j	s_k	$(\xi^{\inf}, \xi^{\inf}, \xi^{\inf})$	$(\xi^1, \xi^{\inf}, \xi^{\inf})$	$(\xi^1, \xi^1, \xi^{\inf})$	$(\xi^1, \xi^{\inf}, \xi^1)$	(ξ^1, ξ^1, ξ^1)
0	0	0	0	0	1	1	1
0	0	1	0	1	1	1	1
0	1	0	0	1	1	1	1
1	0	0	0	0	1	1	1

Since a switch in the vote outcome from 0 to 1 is costly in state $\omega = 0$ and beneficial in state $\omega = 1$, a juror i with *Homo moralis* preferences and degree of morality κ strictly prefers strategy $\xi^{\inf}$ to ξ^1 if and only if the probability that i would either be pivotal or κ-pivotal when the state is $\omega = 0$ exceeds the probability of the same event when the state is $\omega = 1$. Counting in Table 1 the pivotal cells with their associated probabilities in states $\omega = 0$ and 1, as well as the weight attached to them by *Homo moralis*, one finds the following necessary and sufficient condition for the deviation to ξ^1 to be unappealing:

$$\begin{array}{rl} & \mu_0 \cdot [\quad 2p_0^2\,(1-p_0) \quad + \quad p_0^2\,[2\kappa(1-\kappa)+\kappa^2] \quad] \\ > & \mu_1 \cdot [\quad 2p_1\,(1-p_1)^2 \quad + \quad (1-p_1)^2[2\kappa(1-\kappa)+\kappa^2] \quad]. \end{array} \tag{10}$$

In a similar manner, Table 2 lists the signal combinations (s_i, s_j, s_k) for which i is either pivotal or κ-pivotal if (s)he deviates from $\xi^{\inf}$ to ξ^0. Since a switch in the vote outcome from 1 to 0 is costly in state $\omega = 1$ and beneficial in state $\omega = 0$, i strictly prefers strategy $\xi^{\inf}$ to ξ^0 if and only if the probability that i would either be pivotal or κ-pivotal when the state is

$\omega = 1$ exceeds the probability of the same event when the state is $\omega = 0$:

$$\begin{array}{rcl} & \mu_1 \cdot [\quad 2p_1^2(1-p_1) \quad + \quad p_1^2[2\kappa(1-\kappa)+\kappa^2] \quad] \\ > & \mu_0 \cdot [\quad 2p_0(1-p_0)^2 \quad + \quad (1-p_0)^2[2\kappa(1-\kappa)+\kappa^2] \quad]. \end{array} \tag{11}$$

Table 2: Signal realizations (s_i, s_j, s_k) for which a deviation by i from ξ^{inf} to ξ^0 would alter the vote outcome, if: (a) j and k play ξ^{inf}, (b) hypothetically, j were also to play ξ^0 instead of ξ^{inf}, (c) hypothetically, k were also to play ξ^0 instead of ξ^{inf}, (d) hypothetically, both j and k were also to play ξ^0 instead of ξ^{inf}

			(a)	(b)	(c)	(d)
s_i	s_j	s_k	$(\xi^{\text{inf}}, \xi^{\text{inf}}, \xi^{\text{inf}})$	$(\xi^0, \xi^{\text{inf}}, \xi^{\text{inf}})$	$(\xi^0, \xi^0, \xi^{\text{inf}})$	$(\xi^0, \xi^{\text{inf}}, \xi^0)$	(ξ^0, ξ^0, ξ^0)
1	1	1	1	1	0	0	0
1	1	0	1	0	0	0	0
1	0	1	1	0	0	0	0
0	1	1	1	1	0	0	0

Defining the two threshold values

$$\bar{\lambda}^{(\kappa)} \equiv \frac{p_1^2}{(1-p_0)^2} \cdot \frac{\kappa(2-\kappa)+2(1-p_1)}{\kappa(2-\kappa)+2p_0} \tag{12}$$

and

$$\underline{\lambda}^{(\kappa)} \equiv \frac{(1-p_1)^2}{p_0^2} \cdot \frac{\kappa(2-\kappa)+2p_1}{\kappa(2-\kappa)+2(1-p_0)}, \tag{13}$$

we note that the condition derived above for $\boldsymbol{\xi}^{\mathbf{inf}}$ to be a Nash equilibrium in a jury of *Homo oeconomicus* (see (9)) can be written $\underline{\lambda}^{(0)} < \mu_0/\mu_1 < \bar{\lambda}^{(0)}$, and that the condition for $\boldsymbol{\xi}^{\mathbf{inf}}$ to be efficient (see (8)) can be written $\underline{\lambda}^{(1)} < \mu_0/\mu_1 < \bar{\lambda}^{(1)}$.

The reasoning above, together with the observation that a deviation to the reverse strategy ξ^{inv} is clearly dominated by a deviation to either ξ^0 or ξ^1, allows us to state necessary and sufficient conditions for informative voting to be an equilibrium.

Proposition 6 *In a jury consisting of three jurors with Homo moralis preferences with degree of morality κ, for any $\mu_0 \in (0, 1/2)$:*

1. $\boldsymbol{\xi}^{\mathbf{inf}}$ *is a strict Bayesian Nash equilibrium if and only if* $\underline{\lambda}^{(\kappa)} < \frac{\mu_0}{1-\mu_0} < \bar{\lambda}^{(\kappa)}$;

2. $\underline{\lambda}^{(\kappa)}$ *is strictly decreasing and* $\bar{\lambda}^{(\kappa)}$ *is strictly increasing in* κ*, for all* $\kappa \in [0,1]$*;*

3. *if* $\kappa = 1$*,* $\boldsymbol{\xi}^{\mathbf{inf}}$ *is a strict Bayesian Nash equilibrium if and only if* $\boldsymbol{\xi}^{\mathbf{inf}}$ *is efficient.*

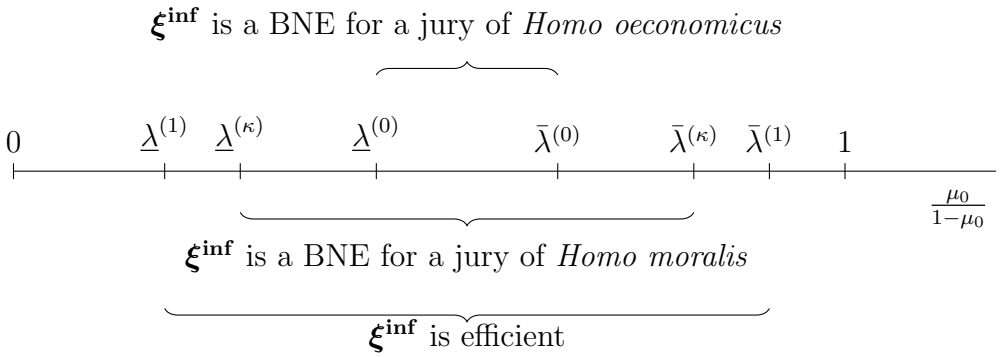

Figure 1: Values of μ_0 for which $\boldsymbol{\xi}^{\mathbf{inf}}$ is a strict Bayesian Nash equilibrium

This result is illustrated in Figure 1, which shows the values of μ_0/μ_1 for which $\boldsymbol{\xi}^{\mathbf{inf}}$ is a strict Bayesian Nash equilibrium, both for a jury composed of three *Homo oeconomicus*, and for a jury composed of three *Homo moralis* with a positive degree of morality $\kappa \in [0,1]$. It also shows the values of μ_0/μ_1 for which $\boldsymbol{\xi}^{\mathbf{inf}}$ is efficient.

In sum, then, *Homo moralis* preferences unambiguously render a small jury more efficient, in the sense that they render informative voting individually rational in settings where such voting is efficient but not individually rational for *Homo oeconomicus*.

Next we examine whether this conclusion also obtains in a large jury.

4.3 A large jury

We restrict attention to the case of an odd number $n = 2m+1$ of voters and check whether the strategy profile $\boldsymbol{\xi}^{\mathbf{inf}}$ whereby all n voters use the informative strategy ξ^{inf} is a (Bayesian) Nash equilibrium for large n.

Proposition 7 *For any fixed values of the parameters* $\mu_0, \mu_1, p_0, p_1, \kappa \in (0,1)$ *such that* $p_1 \geq p_0 > 1/2$*, there exists an integer* N *such that for any jury of size* $n = 2m+1 \geq N$*:*

- *if* $\kappa > 1 - p_0/p_1$*, the informative voting strategy profile* $\boldsymbol{\xi}^{\mathbf{inf}}$ *is a strict Nash equilibrium;*
- *if* $\kappa < 1 - p_0/p_1$ *the informative voting strategy profile* $\boldsymbol{\xi}^{\mathbf{inf}}$ *is not a Nash equilibrium;*

- *if* $\kappa = 1 - p_0/p_1$ *the informative voting strategy profile* $\boldsymbol{\xi}^{\mathbf{inf}}$ *is a strict Nash equilibrium if* $p_0\mu_0 < \mu_1$ *and is not a Nash equilibrium if* $p_0\mu_0 > \mu_1$.

Having thus shown that *Homo moralis* preferences pave the way for informative voting to be sustained as a Nash equilibrium in large juries, as (implicitly) posited by Condorcet, we next ask whether such preferences would also help large juries escape from the use of non-informative voting strategies.

Thus, consider now the situation in which all jurors use a non-informative strategy, for instance ξ^0. Then the decision is 0 independently of the signals received by the jurors. When the jury is large this is clearly a flat Nash equilibrium in a jury composed of jurors with *Homo oeconomicus* preferences, since the outcome would then be unaffected by the deviation of a single juror. The expected payoff at this strategy profile is simply μ_0, the prior probability that the state of nature is 0. We now prove that in a jury composed of jurors with *Homo moralis* preferences, for any strictly positive value of the morality parameter κ each juror strictly prefers to deviate to the informative strategy. While it is still the case that such a deviation by a single juror has no effect on the actual outcome, *Homo moralis* evaluates the outcome under the informative strategy, should a share κ of the other jurors also use it. For a large enough jury, this makes the individual voter κ-pivotal, thus inducing a strict preference ranking over the strategies.

Proposition 8 *For any fixed values of the parameters* $\mu_0, \mu_1, p_0, p_1, \kappa \in (0,1)$ *there exists an integer* N *such that if* $n = 2m + 1 \geq N$*, a strategy profile in which each juror always votes for the same option (0 or 1) independently of her signal, is not a Nash equilibrium.*

Remark that for the result in this proposition to hold there is no condition on the value of κ: for any κ, for n large enough, uninformative voting is not an equlibrium. This was not the case for the previous proposition in which the constraint $\kappa > 1 - p_0/p_1$ appears.

4.4 Conclusion on the large jury problem

It follows from the results reported in Propositions 7 and 8 that the Condorcet Jury Theorem (asymptotic efficient revelation of information) holds for large juries composed of jurors with *Homo moralis* preferences, when the degree of morality κ is large enough. The following theorem sums up these results by characterizing the equilibrium properties of all symmetric strategy profiles (the first two points are taken from the two previous propositions, and we leave to the reader the proof of the third point, which can follow the same lines).

Theorem 2 *For any fixed values of the parameters $\mu_0 \in [0,1]$ and $p_1 \geq p_0 > 1/2$, for any $\kappa \in (0,1]$, if the size of the jury $n = 2m+1$ is large enough:*

- *informative voting, whereby each juror casts a vote for the decision that was suggested to him/her through the private signal (i.e., all jurors use the informative strategy ξ^{inf}), is a strict Nash equilibrium if $\kappa > 1 - p_0/p_1$, and is not a Nash equilibrium if $\kappa < 1 - p_0/p_1$;*
- *uninformative voting, whereby all jurors always vote for the same decision (i.e., all jurors either use strategy ξ^0 or strategy ξ^1)), is not a Nash equilibrium;*
- *all players using the reversed strategy ξ^{inv} is not a Nash equilibrium.*

5 Conclusion

The two problems we studied can be seen as instances of "social dilemmas" but they are different. The first example—the divided majority problem—is a pure multi-person coordination problem. It provides a simple but non-trivial exercise to illustrate how *Homo moralis* preferences can help solve coordination problems (see also Alger and Weibull 2017 [3]). The second example—the rational approach to the Condorcet jury theorem—adds a question of information processing to the coordination issue. This raises the same kind of difficulties that appear in the evolutionary theory of language (Laslier 2003 [21], Demichelis and Weibull 2008 [12], Benz et al. 2011 [6]), and analyzing these issues require the use of Bayesian equilibrium. We find that *Homo moralis* preferences help improve the information aggregation that the jurors can achieve by voting based solely on their private information, without communicating with each other.

In the two problems we studied, the partial morality built into *Homo moralis* preferences impacts the predictions. Importantly, the predictions are often less surprising than those of the standard rational model with materially self-interested individuals: indeed, in the settings we consider *Homo moralis* is often not subject to phenomena often described as "paradoxes" or "curses". This observation is an invitation to add the *Homo moralis* model to the toolkit of political economy for descriptive purposes, and also to deepen the evolutionary analysis of political games.

Appendix

Proof of Proposition 1

If everyone votes sincerely, candidate B wins, and A-supporters achieve utility 0.

1. Consider now an $A2$-supporter who ponders deviating to a vote for $A1$. As a *Homo moralis* with the population of $A2$-supporters as her reference group, when computing her utility for this deviation she evaluates what the outcome would be if, hypothetically, a fraction κ of the n_{A2} supporters of candidate $A2$ would also vote for $A1$. She thus considers what the outcome would be if the number of votes in favor of $A2$ went down from n_{A2} to $(1-\kappa)n_{A2}$, the number of votes for $A1$ went up from n_{A1} to $n_{A1}+\kappa n_{A2}$, and the number of votes for B remained at n_B. This voter would benefit in utility terms from this deviation if and only if the candidate that would win was $A1$ instead of B. The condition for this deviation to be favorable is thus $n_{A1}+\kappa n_{A2} > n_B$, or $\kappa > \kappa^*$.

2. In a similar manner, an $A1$-supporter would strictly prefer to deviate and vote for $A2$ rather than voting sincerely for $A1$ if and only if $n_{A2}+\kappa n_{A1} > n_B$. Since $n_{A2} < n_{A1}$, we have $n_{A2}+\kappa n_{A1} < n_{A1}+\kappa n_{A2}$; in other words, an $A1$-supporter strictly prefers to deviate if and only if an $A2$-supporter also prefers to deviate.

3. In sum, if all other A-supporters vote sincerely, both $A1$- and $A2$-supporters are indifferent between voting sincerely and deviating to some other strategy if $\kappa < \kappa^*$, while $A2$-supporters strictly prefer to deviate if $\kappa > \kappa^*$.

Proof of Proposition 2

Suppose that all A-supporters vote for the same candidate Ak, $k \in \{1,2\}$. Then candidate Ak gets the score $n_{A1}+n_{A2} > n_B$ and thus wins. Now note that:

1. In the *ex post* scenario an $A1$-supporter who considers some deviation, evaluates what the outcome of the vote would be should a proportion κ of her fellow $A1$-supporters also play the deviating strategy. In this hypothetical scenario candidate Ak would get the score $(1-\kappa)n_{A1}+n_{A2}$. The deviation entails a drop in utility if in this hypothetical scenario candidate B wins, i.e., if the score $(1-\kappa)n_{A1}+n_{A2}$ falls short of n_B, the score of candidate B, i.e., if $\kappa > \frac{n_A-n_B}{n_{A1}}$. The deviation has no effect on the deviator's utility if Ak still wins in this hypothetical scenario, i.e., if $\kappa < \frac{n_A-n_B}{n_{A1}}$.

2. Likewise: in the *ex post* scenario an $A2$-supporter who considers some deviation, evaluates what the outcome of the vote would be should a proportion κ of her fellow $A2$-supporters also play the deviating strategy. In this hypothetical scenario candidate Ak would get the score $n_{A1} + (1-\kappa)n_{A2}$. The deviation entails a drop in utility if in this hypothetical scenario candidate B wins, i.e., if the score $n_{A1} + (1-\kappa)n_{A2}$ falls short of n_B, the score of candidate B, i.e., if $\kappa > \frac{n_A - n_B}{n_{A2}}$. The deviation has no effect on the deviator's utility if Ak still wins in this hypothetical scenario, i.e., if $\kappa < \frac{n_A - n_B}{n_{A2}}$.

3. These observations imply the statement in the proposition, since $\frac{n_A - n_B}{n_{A2}} > \frac{n_A - n_B}{n_{A1}}$.

Proof of Proposition 3

If everyone votes sincerely, candidate B wins, and A-supporters achieve utility 0.

1. Consider now an A-supporter who ponders deviating to strategy $\alpha^{(1)}$. In the *ex ante* scenario, this voter evaluates what his expected utility would be if, hypothetically, a share κ of other A-supporters would also use strategy $\alpha^{(1)}$. The score of candidate $A1$ would in this hypothetical scenario be $n_{A1} + \kappa n_{A2}$, that of candidate $A2$ would be $(1-\kappa)n_{A2}$, and that of candidate B would be n_B. Since $n_{A2} < n_B$, it follows that $C^{(\kappa)}(\alpha^{(1)}, \alpha^{\text{si}}) \neq B$ if and only if $C^{(\kappa)}(\alpha^{(1)}, \alpha^{\text{si}}) = A1$, which is true iff $\kappa > (n_B - n_{A1})/n_{A2}$. The expected utility of the deviating A-voter is thus (see (6))

$$U_A^{(\kappa)}(\alpha^{(1)}, \alpha^{\text{si}}) = \begin{cases} 1 + \varepsilon \cdot (n_{A1} - n_{A2})/(n_{A1} + n_{A2}) & \text{if } \kappa > (n_B - n_{A1})/n_{A2} \\ 0 \quad \text{otherwise.} \end{cases} \tag{14}$$

   Hence, $U^{(\kappa)}(\alpha^{(1)}, \alpha^{\text{si}}) > U^{(\kappa)}(\alpha^{\text{si}}, \alpha^{\text{si}})$ if $\kappa > (n_B - n_{A1})/n_{A2}$, while $U^{(\kappa)}(\alpha^{(1)}, \alpha^{\text{si}}) = U^{(\kappa)}(\alpha^{\text{si}}, \alpha^{\text{si}})$ otherwise.

2. Consider now an A-supporter who ponders deviating to strategy $\alpha^{(2)}$. In the *ex ante* scenario, this voter evaluates what his expected utility would be if, hypothetically, a share κ of other A-supporters would also use strategy $\alpha^{(2)}$. The score of candidate $A1$ would in this hypothetical scenario be $(1-\kappa)n_{A1}$, that of candidate $A2$ would be $\kappa n_{A1} + n_{A2}$, and that of candidate B would be n_B. Since $n_{A1} < n_B$, it follows that $C^{(\kappa)}(\alpha^{(2)}, \alpha^{\text{si}}) \neq B$ if and only if $C^{(\kappa)}(\alpha^{(2)}, \alpha^{\text{si}}) = A2$, which is true iff $\kappa > (n_B - n_{A2})/n_{A1}$. The expected utility of the deviating A-voter is thus (see (6))

$$U_A^{(\kappa)}(\alpha^{(2)}, \alpha^{\text{si}}) = \begin{cases} 1 - \varepsilon \cdot (n_{A1} - n_{A2})/(n_{A1} + n_{A2}) & \text{if } \kappa > (n_B - n_{A2})/n_{A1} \\ 0 \quad \text{otherwise.} \end{cases} \tag{15}$$

Hence, $U^{(\kappa)}(\alpha^{(2)}, \alpha^{\text{si}}) > U^{(\kappa)}(\alpha^{\text{si}}, \alpha^{\text{si}})$ if $\kappa > (n_B - n_{A2})/n_{A1}$, while $U^{(\kappa)}(\alpha^{(2)}, \alpha^{\text{si}}) = U^{(\kappa)}(\alpha^{\text{si}}, \alpha^{\text{si}})$ otherwise. But the assumptions in (3) imply $(n_B - n_{A2})/n_{A1} > (n_B - n_{A1})/n_{A2}$, and hence the deviation to strategy $\alpha^{(2)}$ is beneficial only if the deviation to strategy $\alpha^{(1)}$ is beneficial.

3. Combining the two previous paragraphs, this proves that there exists a utility-enhancing deviation if and only if $\kappa > (n_B - n_{A1})/n_{A2}$.

Proof of Proposition 4

Suppose that all A-supporters use strategy $\alpha^{(1)}$. Then candidate $A1$ wins and A-voters have expected utility $1 + \varepsilon \cdot (n_{A1} - n_{A2})/(n_{A1} + n_{A2})$. In the *ex post* scenario an A-supporter who considers some deviation, evaluates what the outcome of the vote would be should a proportion κ of all A-supporters also play the deviating strategy.

1. Consider first a deviation to strategy α^{si}. The hypothetical scores that would obtain should a share κ of other A-voters also use strategy α^{si} instead of $\alpha^{(1)}$ are $n_{A1} + (1-\kappa) \cdot n_{A2}$ for candidate $A1$, $\kappa \cdot n_{A2}$ for candidate $A2$, and n_B for candidate B. Since $\kappa n_{A2} < n_{A1} + (1-\kappa) n_{A2}$, the deviation leads to a drop in expected utility if $n_B > n_{A1} + (1-\kappa) \cdot n_{A2}$, i.e., if $\kappa > \frac{n_A - n_B}{n_{A2}}$, and no change in expected utility otherwise.

2. Consider now a deviation from strategy $\alpha^{(1)}$ to strategy $\alpha^{(2)}$. This deviation leads to the hypothetical scores $(1-\kappa) n_A$ for candidate $A1$, κn_A for candidate $A2$, and n_B for candidate B. The effect on the voter's utility depends on the value of κ:

   - if $\kappa > 1/2$, the deviation implies $C^{(\kappa)}(\alpha^{(2)}, \alpha^{(1)}) \in \{A2, B\}$ and thus a drop in utility (see (6));
   - if $\kappa < 1/2$, there are two cases:
     - if $\kappa > \frac{n_A - n_B}{n_A}$, the deviation implies $C^{(\kappa)}(\alpha^{(2)}, \alpha^{(1)}) = B$ and thus entails a drop in utility;
     - if $\kappa < \frac{n_A - n_B}{n_A}$, the deviation implies $C^{(\kappa)}(\alpha^{(2)}, \alpha^{(1)}) = A1$, leaving the utility unaffected.

3. Since $\frac{n_A - n_B}{n_{A2}} > \frac{n_A - n_B}{n_A}$, we conclude that, if all other A-supporters use strategy $\alpha^{(1)}$, an A-supporter:

   - is indifferent between the strategies $\alpha^{(1)}$, α^{si}, and $\alpha^{(2)}$ if $\kappa < \frac{n_A - n_B}{n_A}$;

- is indifferent between the strategies $\alpha^{(1)}$ and α^{si}, but strictly prefers not to deviate to $\alpha^{(2)}$ if $\frac{n_A-n_B}{n_A} < \kappa < \frac{n_A-n_B}{n_2}$;
- strictly prefers not to deviate from $\alpha^{(2)}$ if $\kappa > \frac{n_A-n_B}{n_{A2}}$.

Proof of Proposition 5

Suppose that all A-supporters use strategy $\alpha^{(2)}$. Then candidate $A2$ wins and A-supporters have expected utility $1-\varepsilon\cdot(n_{A1}-n_{A2})/(n_{A1}+n_{A2})$. In the *ex post* scenario an A-supporter who considers some deviation, evaluates what the outcome of the vote would be should a proportion κ of all A-supporters also play the deviating strategy.

1. Consider first a deviation to strategy α^{si}. The hypothetical scores that would obtain should a share κ of other A-voters also use strategy α^{si} instead of $\alpha^{(2)}$ are $\kappa\cdot n_{A1}$ for candidate $A1$, $(1-\kappa)\cdot n_{A1}+n_{A2}$ for candidate $A2$, and n_B for candidate B. The deviation leads to a drop in expected utility if $n_B > n_{A1}+(1-\kappa)\cdot n_{A2}$, i.e., if $\kappa > \frac{n_A-n_B}{n_{A2}}$, and no change in expected utility otherwise.

2. Consider now a deviation from strategy $\alpha^{(2)}$ to strategy $\alpha^{(1)}$. The hypothetical scores that would obtain with this deviation are κn_A for candidate $A1$, $(1-\kappa)n_A$ for candidate $A2$, and n_B for candidate B. The effect on the voter's expected utility depends on the value of κ as follows:

   - if $\kappa < \min\{1/2, \frac{n_A-n_B}{n_A}\}$, the deviation implies $C^{(\kappa)}(\alpha^{(1)},\alpha^{(2)}) = A2$ and thus no change in expected utility;
   - if $\max\{\kappa, 1-\kappa\}\cdot n_A < n_B$, the deviation implies $C^{(\kappa)}(\alpha^{(1)},\alpha^{(2)}) = B$ and thus a drop in expected utility;
   - if $\kappa > \max\{1/2, \frac{n_B}{n_A}\}$, the deviation implies $C^{(\kappa)}(\alpha^{(1)},\alpha^{(2)}) = A1$ and thus an increase in expected utility.

3. Noting that $\frac{n_A-n_B}{n_{A2}} > \frac{n_A-n_B}{n_A}$, and that $\frac{n_A-n_B}{n_A} < \frac{n_B}{n_A} \Leftrightarrow \frac{n_A-n_B}{n_A} < \frac{1}{2} < \frac{n_B}{n_A} \Leftrightarrow n_A < \frac{2}{3}$, which is true by assumption (see (3)), we can conclude that, if all other A-supporters use strategy $\alpha^{(2)}$, an A-supporter:

   - is indifferent between the strategies $\alpha^{(2)}$, α^{si}, and $\alpha^{(1)}$ if $\kappa < \frac{n_A-n_B}{n_A}$;
   - is indifferent between the strategies $\alpha^{(2)}$ and α^{si}, but strictly prefers not to deviate to $\alpha^{(1)}$ if $\frac{n_A-n_B}{n_A} < \kappa < \frac{n_A-n_B}{n_{A2}}$;
   - strictly prefers not to deviate from $\alpha^{(2)}$ if $\frac{n_A-n_B}{n_{A2}} < \kappa < \frac{n_B}{n_A}$;

- strictly prefers to deviate from $\alpha^{(2)}$ to $\alpha^{(1)}$ if $\kappa > \frac{n_B}{n_A}$.

Proof of Proposition 7

Let $\pi(\xi_i, \boldsymbol{\xi_{-i}})$ denote i's expected material payoff if i plays strategy ξ_i and the other jurors play strategy profile $\boldsymbol{\xi_{-i}} \in X^{2m}$. Writing $A_\omega(\xi_i, \boldsymbol{\xi_{-i}})$ for the probability that the decision is correct in state ω, we thus have:

$$\pi(\xi_i, \boldsymbol{\xi_{-i}}) = \mu_0 A_0(\xi_i, \boldsymbol{\xi_{-i}}) + \mu_1 A_1(\xi_i, \boldsymbol{\xi_{-i}}). \tag{16}$$

Now, under majority rule the probability that the right decision is taken in state ω equals the probability that at least $m+1$ jurors vote ω when the state is ω. Using the following notation for the binomial probability of t or more successes out of T:

$$B^+(p, t, T) = \sum_{k=t}^{T} C_T^k p^k (1-p)^{T-k}, \tag{17}$$

we immediately obtain that if all the jurors vote informatively, i.e., if $(\xi_i, \boldsymbol{\xi_{-i}}) = (\xi^{\text{inf}}, \boldsymbol{\xi^{\text{inf}}})$, the probability that the decision is correct in state ω equals the probability that at least $m+1$ jurors receive the signal ω. In other words,

$$\pi(\xi^{\text{inf}}, \boldsymbol{\xi^{\text{inf}}}) = \mu_0 A_0(\xi^{\text{inf}}, \boldsymbol{\xi^{\text{inf}}}) + \mu_1 A_1(\xi^{\text{inf}}, \boldsymbol{\xi^{\text{inf}}}), \tag{18}$$

where

$$A_0(\xi^{\text{inf}}, \boldsymbol{\xi^{\text{inf}}}) = B^+(p_0, m+1, 2m+1) \tag{19}$$

and

$$A_1(\xi^{\text{inf}}, \boldsymbol{\xi^{\text{inf}}}) = B^+(p_1, m+1, 2m+1). \tag{20}$$

We now derive conditions required for a juror, say i, to prefer playing ξ^{inf} to deviating to strategy ξ^0. If this juror has *Homo moralis* preferences with degree of morality $\kappa \in [0,1]$, (s)he evaluates the consequences of the deviation on his/her expected material payoff, should each other juror play ξ^{inf} with probability $1-\kappa$ and ξ^0 with probability κ. Taking into account this reasoning, let $v_0^{(\kappa)}$ be the number of votes 0 and $v_1^{(\kappa)}$ be the number of votes 1 that *Homo moralis* envisages. Because i votes 0, the probability of a correct decision, based

on the reasoning of *Homo moralis* with degree of morality κ, is:

$$\begin{aligned} \pi^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}) &= \Pr[v_0^{(\kappa)} \geq m \text{ and } \omega = 0] + \Pr[v_1^{(\kappa)} \geq m+1 \text{ and } \omega = 1] \\ &= \mu_0 A_0^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}) + \mu_1 A_1^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}), \end{aligned} \tag{21}$$

where $A_\omega^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}})$ is the probability that at least $m+1$ jurors vote ω in state ω, taking into account the hypothetical scenario that each other juror play ξ^{inf} with probability $1-\kappa$ and ξ^0 with probability κ. Formally:

$$A_0^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}) = \Pr\left[v_0 \geq m \mid \omega = 0\right] = B^+(p_0', m, 2m) \tag{22}$$

$$A_1^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}) = \Pr\left[v_1 \geq m+1 \mid \omega = 1\right] = B^+(p_1', m+1, 2m), \tag{23}$$

where

$$p_0' = (1-\kappa)p_0 + \kappa \tag{24}$$

and

$$p_1' = (1-\kappa)p_1. \tag{25}$$

Now, note that as $m \to \infty$, both $B^+(p, m+1, 2m+1)$ and $B^+(p, m, 2m+1)$ tend either to 0 or to 1 depending on whether p is smaller or greater than $1/2$. Specifically:

- both $A_0(\xi^{\text{inf}}, \boldsymbol{\xi}^{\mathbf{inf}})$ and $A_1(\xi^{\text{inf}}, \boldsymbol{\xi}^{\mathbf{inf}})$ are increasing in m and tend to 1 as $m \to \infty$, since (by assumption) both $p_0 > 1/2$ and $p_1 > 1/2$;
- $A_0^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}})$ is increasing in m and tends to 1 as $m \to \infty$, since $p_0' = (1-\kappa)p_0 + \kappa > p_0 > 1/2$;
- $A_1^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}})$ is increasing in m and tends to 1 as $m \to \infty$ if $p_1' = (1-\kappa)p_1 > 1/2$, but is is decreasing in m and tends to 0 as $m \to \infty$ if $p_1' = (1-\kappa)p_1 < 1/2$.

Taken together, these observations imply the following:

- $\pi(\xi^{\text{inf}}, \boldsymbol{\xi}^{\mathbf{inf}})$ (see (18)) is increasing in m and tends to $\mu_0 + \mu_1$ as $m \to \infty$;
- $\pi(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}})$ is increasing in m and tends to $\mu_0 + \mu_1$ as $m \to \infty$ if $(1-\kappa)p_1 > 1/2$;
- $\pi(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}})$ tends to μ_0 as $m \to \infty$ if $(1-\kappa)p_1 < 1/2$.

Clearly, the deviation to ξ^0 is thus not viable if $(1-\kappa)p_1 < 1/2$ and m is large enough. We take note of this result:

Lemma 1 *Suppose that $\kappa > 1 - 1/(2p_1)$. If m is large enough, each juror strictly prefers to play $\xi^{\inf}$ than to deviate to ξ^0.*

By contrast, if $\kappa < 1 - 1/(2p_1)$, both $\pi(\xi^{\inf}, \boldsymbol{\xi}^{\mathbf{inf}})$ and $\pi(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}})$ tend to 1 as $m \to \infty$, hence it is a priori impossible to determine whether a deviation to ξ^0 is viable. In order to go further, we remark that the kind of binomial sums that appear in the calculations can be approximated using the large deviation theory for binomial laws; see for instance Arratia and Gordon (1989 [4]).

Let x_n and y_n, $n \in \mathbb{N}$ be two sequences of real numbers. We say that x and y are equivalent, and we write $x \sim y$, when $\lim_{n\to\infty} x_n/y_n = 1$. Likewise, we use the notation $x \ll y$ when $\lim_{n\to\infty} x_n/y_n = 0$.

Lemma 2 *Let a and q be two real numbers such that $0 < q < a < 1$, and*

$$\begin{aligned} H = H(q,a) &= a \log \frac{a}{q} + (1-a) \log \frac{1-a}{1-q}, \\ r = r(q,a) &= \frac{q}{1-q} \Big/ \frac{a}{1-a} = \frac{q(1-a)}{a(1-q)}, \\ \rho = \rho(q,a) &= \frac{1}{(1-r)\sqrt{2\pi a(1-a)}}. \end{aligned}$$

Then, when n tends to infinity:

$$B^+(q, an, n) \sim \rho e^{-Hn}.$$

We apply these formula to the expressions above, in order to check whether the deviation is profitable. Recall that we defined $A_0(\xi^{\inf}, \boldsymbol{\xi}^{\mathbf{inf}}) = B^+(p_0, m+1, 2m+1)$, so that $1 - A_0(\xi^{\inf}, \boldsymbol{\xi}^{\mathbf{inf}}) = B^+(q_0, m+1, 2m+1)$, where $q_0 = 1 - p_0$. We likewise re-write:

$$\begin{aligned} 1 - A_0(\xi^{\inf}, \boldsymbol{\xi}^{\mathbf{inf}}) &= B^+(q_0, m+1, 2m+1) \\ 1 - A_1(\xi^{\inf}, \boldsymbol{\xi}^{\mathbf{inf}}) &= B^+(q_1, m+1, 2m+1) \\ 1 - A_0^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}) &= B^+(q_0', m, 2m) \\ 1 - A_1^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}) &= B^+(q_1', m+1, 2m), \end{aligned}$$

where

$$q_0 = 1 - p_0$$
$$q_1 = 1 - p_1$$
$$q_0' = (1-\kappa)q_0$$
$$q_1' = (1-\kappa)q_1 + \kappa.$$

Writing the expected utility gain from deviating in the form:

$$\Delta^0 = \pi^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}) - \pi(\xi^{\inf}, \boldsymbol{\xi}^{\mathbf{inf}}),$$

we have

$$\Delta^0 = \mu_0 A_0^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}) + \mu_1 A_1^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}}) - \mu_0 A_0(\xi^{\inf}, \boldsymbol{\xi}^{\mathbf{inf}}) - \mu_1 A_1(\xi^{\inf}, \boldsymbol{\xi}^{\mathbf{inf}}), \quad (26)$$

which can also be written

$$\begin{aligned}\Delta^0 &= \mu_0 \cdot \left[B^+(q_0, m+1, 2m+1) - B^+(q_0', m, 2m)\right] \\ &\quad +\mu_1 \cdot \left[B^+(q_1, m+1, 2m+1) - B^+(q_1', m+1, 2m)\right].\end{aligned} \quad (27)$$

Let us now apply Lemma 2 to the first of these four binomial sums. Let $n = 2m+1$ and $a = 1/2$, then $an = m + 1/2$ so that having an or more Bernouilli successes means having $m+1$ or more. Because $q_0 < 1/2$ the lemma applies and writes: $B^+(q_0, m+1, 2m+1) \sim \rho(q_0, 1/2)e^{-H(q_0,1/2)n}$. Note that $H(q_0, 1/2) = (1/2)\log[4q_0(1-q_0)]$ so that $e^{-H(q_0,1/2)n} = [4q_0(1-q_0)]^{-n/2}$ and we obtain:

$$B^+(q_0, m+1, 2m+1) \sim \rho(q_0, 1/2)[4q_0(1-q_0)]^{-(2m+1)/2}.$$

The same reasoning works for the other sums in (27), because q_0, q_0', q_1, and q_1' are all strictly smaller than $1/2$. In each case, i.e., for $q \in \{q_0, q_1, q_0', q_1'\}$, the same short computation gives the same form for H:

$$H(q, 1/2) = -(1/2)\log[4(1-q)q]$$

and we obtain:

$$B^+(q_0, m+1, 2m+1) \sim \rho(q_0, 1/2)\left[4q_0(1-q_0)\right]^{(2m+1)/2}$$

$$B^+(q_1, m+1, 2m+1) \sim \rho(q_1, 1/2)\left[4q_1(1-q_1)\right]^{(2m+1)/2}.$$

$$B^+(q_0', m, 2m) \sim \rho(q_0', 1/2) \left[4q_0'(1-q_0')\right]^m$$
$$B^+(q_1', m+1, 2m) \sim \rho(q_0', 1/2) \left[4q_1'(1-q_1')\right]^m.$$

Having in mind the shape of the function $x \mapsto x(1-x)$, one can see that, because $q_0' = (1-\kappa)q_0 < q_0 < 1/2$, $q_0'(1-q_0') < q_0(1-q_0)$, so that the term $1 - A_0^{(\kappa)}(\xi^0, \boldsymbol{\xi}^{\mathbf{inf}})$ tends to zero faster than the term $1 - A_0(\xi^{\inf}, \boldsymbol{\xi}^{\mathbf{inf}})$. Hence, when m tends to infinity:

$$B^+(q_0', m, 2m) \ll B^+(q_0, m+1, 2m+1).$$

We thus reach the following conclusion for the first term in (27): for m large enough,

$$\mu_0 \cdot [B^+(q_0, m+1, 2m+1) - B^+(q_0', m, 2m)] \sim \mu_0 \cdot B^+(q_0, m+1, 2m+1).$$

Turning now to the second term in (27), note that $q_1 < q_1' < 1/2$, so that $q_1'(1-q_1') > q_1(1-q_1)$. It follows that

$$B^+(q_1, m+1, 2m+1) \ll B^+(q_1', m+1, 2m).$$

We thus reach the following conclusion for the second term in (27):

$$\mu_1 \cdot [B^+(q_1, m+1, 2m+1) - B^+(q_1', m+1, 2m)] \sim -\mu_1 \cdot B^+(q_1', m+1, 2m).$$

Taken together, these observations imply that when m tends to infinity,

$$\Delta^0 \sim \mu_0 \cdot B^+(q_0, m+1, 2m+1) - \mu_1 \cdot B^+(q_1', m+1, 2m). \tag{28}$$

Recalling that $q_0 < 1/2$ and $q_1' < 1/2$, we have

$$q_0(1-q_0) < q_1'(1-q_1') \iff q_1' > q_0 \iff p_0 > p_1(1-\kappa).$$

Hence:

- $p_0 > p_1(1-\kappa) \implies B^+(q_0, m+1, 2m+1) \ll B^+(q_1', m+1, 2m)$,
- $p_0 < p_1(1-\kappa) \implies B^+(q_1', m+1, 2m) \ll B^+(q_0, m+1, 2m+1)$,

and it follows that:

- If $\kappa > 1 - p_0/p_1$, then $\Delta^0 \sim -\mu_1 \cdot B^+(q_1', m+1, 2m)$.

- If $\kappa < 1 - p_0/p_1$, then $\Delta^0 \sim \mu_0 \cdot B^+(q_0, m+1, 2m+1)$.

From this we can infer the sign of Δ^0 when m is large: for instance in the first case, because the ratio $\Delta^0/(\mu_1 \cdot B^+(q_1', m+1, 2m))$ tends to -1 when m tends to infinity, there exists $\bar{m}$ such that for all $m \geq \bar{m}$, $\Delta^0 < 0$. With this reasoning we complement in a unique statement Lemma 1 and conclude:

- If $\kappa > 1 - p_0/p_1$, then $\Delta^0 < 0$ for m large, in which case deviating to ξ^0 is not profitable;
- If $\kappa < 1 - p_0/p_1$, then $\Delta^0 > 0$ for m large, in which case deviating to ξ^0 is profitable.

In the knife-edge case $\kappa = 1 - p_0/p_1$,that is $q_0 = q_1'$, equation (28) writes:

$$\Delta^0 \sim \mu_0 B^+(q_0, m+1, 2m+1) - \mu_1 B^+(q_0, m+1, 2m).$$

Decomposing the binomial sums and re-arranging the terms, one obtains:

$$\begin{aligned}
\Delta^0 &\sim \mu_0 \sum_{k=m+1}^{2m+1} C_{2m+1}^k q_0^k (1-q_0)^{2m+1-k} - \mu_1 \sum_{k=m+1}^{2m} C_{2m+1}^k q_0^k (1-q_0)^{2m-k} \\
&= \sum_{k=m+1}^{2m} \left[\mu_0(1-q_0) - \mu_1\right] C_{2m+1}^k q_0^k (1-q_0)^{2m-k} + \mu_0 q_0^{2m+1} \\
&= \left[\mu_0(1-q_0) - \mu_1\right] B^+(q_0, m+1, 2m) + \mu_0 q_0^{2m+1}.
\end{aligned}$$

Because $B^+(q_0, m+1, 2m)$ decreases at rate $[4q_0(1-q_0)]^{-m}$, the last term ($\mu_0 q_0^{2m+1}$) is negligible, so that

$$\Delta^0 \sim \left[\mu_0(1-q_0) - \mu_1\right] B^+(q_0, m+1, 2m),$$

implying that Δ^0 has the same sign as $\mu_0(1-q_0) - \mu_1$. We conclude that in the knife-edge case, ξ^0 destabilizes informative voting if and only if $\mu_0(1-q_0) > \mu_1$ (when m tends to infinity).

The symmetric conclusions are reached for a deviation from $\xi^{\inf}$ to ξ^1, the threshold value for κ being equal to $1 - p_1/p_0$ instead of $1 - p_0/p_1$. Since κ is positive only one of these threshold values is relevant, however:

- if $p_0 \geq p_1$, then $1 - p_0/p_1 \leq 0$, and a deviation to ξ^0 is not profitable for any $\kappa \in [0,1]$; but if $p_0 > p_1$ the threshold value $1 - p_1/p_0$ is positive, implying that a deviation to ξ^1 is profitable if $\kappa \in [0, 1 - p_1/p_0)$;

- as shown above, the opposite conclusion holds if $p_1 \geq p_0$.

Putting all this together we find the statement in the proposition, where for simplicity and without loss of generality, we only treat the case $p_1 \geq p_0$.

Proof of Proposition 8

Suppose that all jurors use strategy ξ^0. The probability that the decision is correct is then:

$$\pi(\xi^0, \boldsymbol{\xi^0}) = \mu_0. \tag{29}$$

Note that this is also the utility of a *Homo moralis* with any degree of morality κ who uses strategy ξ^0 given that all other jurors do so as well.

We now derive conditions required for a *Homo moralis* juror i to prefer to deviate from ξ^0 to the informative strategy ξ^{inf}. Such a juror evaluates the consequences of the deviation on his/her expected material payoff, should each other juror play ξ^0 with probability $1-\kappa$ and ξ^{inf} with probability κ. As above, let $v_0^{(\kappa)}$ be the number of votes 0 and $v_1^{(\kappa)}$ be the number of votes 1 that *Homo moralis* envisages. Because i votes informatively, the probability of a correct decision, based on the reasoning of *Homo moralis* with degree of morality κ, is:

$$\pi^{(\kappa)}(\xi^{\text{inf}}, \boldsymbol{\xi^0}) = \quad \mu_0 \cdot \left[p_0 B^+(1-\kappa q_0, m, 2m) + q_0 B^+(1-\kappa q_0, m+1, 2m)\right] \tag{30}$$

$$+ \quad \mu_1 \cdot \left[p_1 B^+(\kappa p_1, m, 2m) + q_1 B^+(\kappa p_1, m+1, 2m)\right]. \tag{31}$$

In this expression line (30) corresponds to the probability of a correct decision in state of Nature $\omega = 0$. The decision is correct in this state if either i receives signal 0 and at least m other voters vote 0, or i receives the wrong signal 1 and at least $m+1$ others vote 0. Any other voter votes 0 either if (s)he uses strategy ξ^0 (a scenario to which *Homo moralis* attaches weight $1-\kappa$), or when (s)he uses strategy ξ^{inf} (a scenario to which *Homo moralis* attaches weight κ), and receives signal 0; so in state $\omega = 0$ *Homo moralis* ponders a hypothetical world in which each other voter has probability $1-\kappa+\kappa p_0 = 1-\kappa q_0$ to vote 0. Line (31) likewise counts the votes 1 in state ω_1, given that *Homo moralis* ponders a hypothetical world in which each other voter uses strategy ξ^0 with probability $1-\kappa$ and strategy ξ^{inf} with probability κ), and thus votes 1 with probability κp_1.

Since $q_0 < 1/2$ (by assumption), we have $1-\kappa q_0 > 1/2$, which implies that both binomial sums in line (30) tend to 1. Hence, line (30) tends to μ_0.

In line (31), as $m \to \infty$ both binomial sums tend to 1 if $\kappa p_1 > 1/2$ and to 0 if $\kappa p_1 < 1/2$.

There are thus two cases:

- If $\kappa p_1 > 1/2$: the expected utility of the deviation to $\xi^{\inf}$ tends to $\mu_0 + \mu_1$, and since this strictly exceeds μ_0 the deviation is strictly profitable (for m large enough).
- If $\kappa p_1 < 1/2$: the expected utility of the deviation tends to μ_0, so the benefit from deviating tends to 0. In order to check its sign we write the expected benefit as follows:

$$\begin{aligned}\Delta = \quad & \pi^{(\kappa)}(\xi^{\inf}, \boldsymbol{\xi^0}) - \pi^{(\kappa)}(\xi^0, \boldsymbol{\xi^0}) \\ = \quad & \mu_0 \cdot \left[p_0(B^+(1-\kappa q_0, m, 2m) - 1) + q_0(B^+(1-\kappa q_0, m+1, 2m) - 1)\right] \\ + \quad & \mu_1 \cdot \left[p_1 B^+(\kappa p_1, m, 2m) + q_1 B^+(\kappa p_1, m+1, 2m)\right].\end{aligned}$$

  Since $B^+(p,t,T) - 1 = -B^+(1-p,t,T)$, this can in turn be written as follows:

$$\begin{aligned}\Delta = \quad - \quad & \mu_0 \cdot \left[p_0 B^+(\kappa q_0, m+1, 2m) + q_0 B^+(\kappa q_0, m, 2m)\right] \\ + \quad & \mu_1 \cdot \left[p_1 B^+(\kappa p_1, m, 2m) + q_1 B^+(\kappa p_1, m+1, 2m)\right].\end{aligned}$$

  Recalling that $q_0 < 1/2 < p_1$, so that $\kappa q_0 < \kappa p_1 < 1/2$, we note that the term in the first square brackets goes to 0 faster than the second one (see Lemma 2 in the proof of Proposition 7). Hence, for m large enough, Δ is positive and deviating is strictly profitable.

The knife-edge case $\kappa p_1 = 1/2$ is easily solved by writing the term in square brackets in line (31) as follows:

$$\begin{aligned} & p_1 B^+(\kappa p_1, m, 2m) + q_1 B^+(\kappa p_1, m+1, 2m) \\ = \quad & p_1 B^+(1/2, m, 2m) + q_1 B^+(1/2, m+1, 2m) \\ > \quad & q_1 \left[B^+(1/2, m, 2m) + B^+(1/2, m+1, 2m)\right] \\ = \quad & q_1,\end{aligned}$$

which proves that, in this case, $\pi^{(\kappa)}(\xi^{\inf}, \boldsymbol{\xi^0}) > \mu_0 + q_1\mu_1 > \mu_0$, making the deviation profitable.

The same reasoning can be applied to show that, for m large enough, a deviation from ξ^1 to $\xi^{\inf}$ is profitable for any $\kappa > 0$. This completes the proof.

References

[1] Ingela Alger and Jean-François Laslier (2020) "Homo moralis goes to the voting booth: a new theory of voter turnout" Working Paper Paris and Toulouse Schools of Economics.

[2] Ingela Alger and Jörgen Weibull (2013) "Homo moralis: preference evolution under incomplete information and assortative matching" *Econometrica* 81: 2269—2302.

[3] Ingela Alger and Jörgen Weibull (2017) "Strategic behavior of moralists and altruists" *Games* 8(3): 38.

[4] Richard Arratia and Louis Gordon (1989) "Tutorial on large deviations for the binomial distribution" *Bulletin of Mathematical Biology* 51: 15—131.

[5] David Austen-Smith and Jeffrey S. Banks (1996): "Information aggregation, rationality, and the Condorcet jury theorem", *American Political Science Review* 90: 34—45.

[6] Anton Benz, Christian Ebert, Gerhard Jäger, and Robert van Rooil (eds.) (2011) *Language, Games, and Evolution: Trends in Current Research on Language and Game Theory*. Berlin, Heidelberg: Springer.

[7] Carisa L. Bergner and Peter K. Hatemi (2017) "Integrating genenetics into the study of electoral behavior" pp.367—405 in: Kai Arzheimer, Jocelyn Evans and Michael lewis-Beck (eds.) *The Sage Handbook of Electoral Behavior. Volume 1*. London: Sage.

[8] Ken Binmore (1994) *Playing fair: Game theory and the social contract*. Cambridge, Mass. MIT Press.

[9] Condorcet (1785) *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Paris: Imprimerie Royale.

[10] Gary W. Cox (1997) *Making Votes Count: Strategic coordination in the World's Electoral Systems*. Cambridge: Cambridge University Press.

[11] Anthony Downs (1957) *An Economic Theory of Democracy*. New York: Harper and Row.

[12] Demichelis, Stefano, and Jorgen W. Weibull. 2008. "Language, meaning, and games: a model of communication, coordination, and evolution" *American Economic Review*, 98(4): 1292—1311.

[13] Timothy Feddersen and Wolfang Pesendorfer (1997) "Voting Behavior and Information Aggregation in Elections with Private and Common Values" *Econometrica* 65: 1029—1058.

[14] Timothy Feddersen and Wolfang Pesendorfer (1998) "Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting" *American Political Science Review* 92: 23—35.

[15] Robert Forsythe, Roger Myerson, Thomas Rietz, and Robert Weber. 1993. An experiment on coordination in multi-candidate elections: The importance of polls and election histories. *Social Choice and Welfare*, 10(3), pp.223—247.

[16] Dino Gerardi and Leeat Yariv (2008): "Information acquisition in committees", *Games and Economic Behavior* 62: 436—459.

[17] Hans Gersbach and Volker Hahn (2008): "Should the invidual voting records of central bankers be published?", *Social Choice and Welfare* 30: 655—683.

[18] Donald P. Green and Ian Shapiro (1994) Pathologies of Rational Choice Theory, *A Critique of Applications in Political Science*, New Haven, Conn.: Yale University Press.

[19] Immanuel Kant (1785) *Grundlegung zur Metaphysik der Sitten*. Trad: Mary Gregor and Jens Timmermann (2011) *Groundwork of the Metaphysics of Morals: A German-English Edition*. Cambridge, Mass.: Cambridge University Press.

[20] Yukio Koriyama and Balázs Szentes (2009): "A resurrection of the Condorcet jury theorem", *Theoretical Economics* 4: 227—252.

[21] Jean-François Laslier (2003) "The evolutionary analysis of signal games" in *Cognitive Economics*, edited by P. Bourgine and J.-P. Nadal, Heidelberg : Springer, pp. 281—291.

[22] Jean-François Laslier and Jörgen Weibull (2013): "An incentive-compatible Condorcet jury theorem", *The Scandinavian Journal of Economics* 115(1), 84—-108.

[23] Jacques Lesourne, André Orléan, and Bernard Walliser (eds.) (2006) *Evolutionary Microeconomics*. Berlin: Springer.

[24] John Maynard Smith (1982) *Evolution and the Theory of Games*. Cambridge, Mass.: Cambridge Universiy Press.

[25] Hervé Moulin (1995) *Cooperative Microeconomics: A Game-Theoretical Introduction*. Cambridge, Mass.: Princeton Universiy Press.

[26] Roger Myerson and Robert Weber (1993) "A Theory of Voting Equilibria" *American Political Science Review*, 87, pp. 102–114.

[27] Roger Myerson (2002) "Comparison of Scoring Rules in Poisson Voting Games" *Journal of Economic Theory*, 103, pp. 219—251.

[28] Roger Myerson and Jörgen Weibull (2015) "Tenable strategies and settled equilibria" *Econometrica*, 83(3), pp. 943—976.

[29] Martin A. Nowak and Karl Sigmund (2005) "Evolution of indirect reciprocity" *Nature* 437: 1293—1295.

[30] Elinor Ostrom (1998) "A behavioral approach to the rational choice theory of collective action: Presidential Address, American Political Science Association, 1997" *American Political Science Review* 92(1): 1—22.

[31] Michael B. Petersen (2015) "Evolutionary political psychology" in: D. Buss (Ed.) *Handbook of Evolutionary Psychology*. Vol. 2. p. 1084-1102. John Wiley.

[32] John Roemer (2019) *How We Cooperate*. Yale University Press.

[33] Jean-Jacques Rousseau (1755) *Discours sur l'origine et les fondements de l'inégalité parmi les hommes*. Reprinted in : *Ecrits politiques*, 1992, Le livre de Poche.

[34] Glendon Shubert (1982) "Evolutionary Politics" *Western Political Quarterly* 175—193.

[35] Thomas D. Seeley (2010) *Honeybee Democracy*. Princeton, NJ: Princeton University Press.

[36] Jim Sidanus and Robert Kurzban (2013) "Toward an evolutionary informed political psychology" in: L. Huddy, D. O. Sears, and J. S. Levy *The Oxford Handbook of Political Psychology* pp. 205—236. Oxford University Press.

[37] Laura B. Stephenson, John H. Aldrich, and André Blais (2018) *The Many Faces of Strategic Voting: Tactical Behavior in Electoral Systems Around the World*. University of Michigan Press.

[38] David J. T. Sumpter (2010) *Collective Animal Behavior*. Princeton, NJ: Princeton University Press.

[39] Franz de Waal (1996) *Good Natured: The origins of right and wrong in in humans and other animals*. Cambridge, Mass.: Cambridge University Press.

[40] Reena H. Walker, Andrew J. King, J. Weldon McNutt, and Neil R. Jordan (2017) "Sneeze to leave: African wild dogs (Lycaon pictus) use variable quorum thresholds facilitated by sneezes in collective decisions" *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 284. DOI: 10.1098/rspb.2017.0347.