Online Appendix for
'Birds of a Feather … Enforce Social Norms? Interactions among Culture, Norms, and Strategy'

Hongyi Li and Eric Van den Steen

February 14, 2021

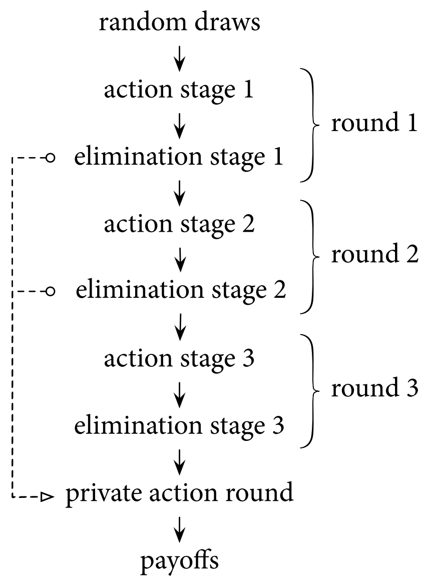
This online appendix discusses details of the formal model in Section 6 of the paper. It aims to provide a self-contained description of the formal analysis. This means that the following discussion overlaps substantially with that of Section 6 – but the discussion also complements the sometimes-informal exposition of Section 6 by filling in some details of the equilibrium, making some features (such as the model timeline) more explicit, and providing details of the proof of Proposition 1.

The paper's formal model is described in Appendix A. The 'social norm equilibrium' is specified in Appendix B. Proposition 1b is restated, a proof is outlined, and examples of the incentive constraints induced by the 'social norm equilibrium' are provided in Appendix C. Finally, we consider Proposition 1a in Appendix D.

A Model and Timeline

There is an organization with $N = 5$ positions to be filled. The employee in position n will choose a public action $a_n \in \{L, R\}$ and a private action $b_n \in \{L, R\}$. (Hereafter, we simply call a_n an *action*; b_n, a *private action.*) Employees are drawn randomly from an infinite pool with two *kinds* of employees, in equal proportion: those who prefer action L and those who prefer action R. Let $\tau_n \in \{L, R\}$ denote the kind of employee in position n. Amongst each kind, proportion $1 - \gamma$ are 'regular' types, labeled l and r, whereas proportion γ are 'strong' types, labeled L and R. The relative proportions of the four employee types (L, l, r, R) are thus $\left(\frac{\gamma}{2}, \frac{1-\gamma}{2}, \frac{1-\gamma}{2}, \frac{\gamma}{2}\right)$. Our analysis will focus on the case of vanishingly rare strong types, $\gamma \to 0$.

The game consists of two parts. In Part I, after employees are randomly drawn, up to three rounds are played; each round consists of an action stage followed by an elimination stage. In each action stage of Part I, every employee n simultaneously chooses (or re-chooses, in the second and third rounds) an action $a_n \in \{L, R\}$. In each elimination stage of Part I, each employee simultaneously decides which other employees to vote against; any employee receiving a majority ($> N/2$) of 'against' votes is immediately eliminated and replaced with a new random draw from the pool. Each employee who is not eliminated in a given round incurs an 'enforcement' cost of $\epsilon > 0$ for each employee who is eliminated in that round. At the end of the third elimination stage, the game proceeds to Part II. Alternatively, after any elimination stage, if nobody voted against anybody else in that round, then the game proceeds immediately to Part II. In Part II, every employee n simultaneously chooses a private action b_n. All employees' action and voting choices are public. To sum up, the timeline of the game is:



The dotted-line arrows highlight – as mentioned above – that at the end of a given elimination stage, if nobody voted against anyone else in that stage, then the game skips ahead to the private action round.

For the second and third rounds of Part I, the action choice a_n for each position replaces the corresponding action choice from the previous round; so, only the last-played action stage of Part I counts. If an employee n is eliminated (and replaced) in the third round, his action choice a_n from the third action stage is retained; his replacement does not get to re-choose a_n, but gets to choose b_n in the private action round of Part II.

At the end of the game, final payoffs are realized. The regular employee in position n at the end of the game receives utility

$$U_n = U + (V J_{a_n=\tau_n} + v J_{b_n=\tau_n}) + \sum_m (W J_{a_m=\tau_n} + w J_{b_m=\tau_n}) - k_n \epsilon$$

where $J_X = 1$ if condition X is satisfied and equals -1 if it is not; and where k_n is the number of other employees that were eliminated in rounds where employee n was present and survived. Here, $U > 0$ is some benefit from being part of the organization, $V, v > 0$ are the benefits/costs from the employee's own choices, and $W, w > 0$ are benefits/costs that an employee enjoys/suffers from the choices of others in the organization. For a regular employee i who is eliminated before the game ends, her final payoff is $-k_i \epsilon$, where k_i is the number of other employees that were eliminated in rounds where i was present and survived.

Strong-type employees (types L and R) choose mechanically. We specify their strategies in Appendix B. Essentially, they always take their preferred action, and vote against employees who they believe are sufficiently likely to be of the opposite kind.

B The 'Social Norm Equilibrium'

To construct the 'social norm equilibrium', we start with some terminology, turn to describe the strategies, then specify the belief-updating rule.

Terminology Here, we reproduce – from Section 6 of the paper – the notation for perceived beliefs (e.g., $l/r/R$), the concepts of L- and R-norms and L- and R-majorities; and associated notation. The reader who is already familiar with Section 6 may skip ahead to the 'Strategies' heading.

Recall that a new random draw from the pool is a mixture of types $L/l/r/R$ with 'prior' probabilities $\frac{\gamma}{2}/\frac{1-\gamma}{2}/\frac{1-\gamma}{2}/\frac{\gamma}{2}$. At any decision node, we call public beliefs about an employee's type the *perceived type* of that employee. At any information set, a perceived type always corresponds to some nonempty subset of the four types with probabilities in relative proportion to the prior (e.g., l/r in relative proportions $\frac{1}{2}/\frac{1}{2}$).

There are only three kinds of perceived types that occur in the 'social norm equilibrium', on- or off-path. For perceived types l, L, and l/L, i.e., the employee is of type l or L with probability one, we say the employee is *L-revealed.* Analogously, for perceived types r, R, and r/R, we say the employee is *R-revealed.* Finally, for perceived types that include both l and r, say $L/l/r$ or l/r, we say the employee is *unrevealed.* Notice that for small γ, unrevealed employees are roughly equally likely to be either type-l or -r.

We say that an employee *acts like kind L* if she chooses action L. She *votes like kind L* if she votes against, and only against, R-revealed employees. (Acting and voting like type R is analogous.) An employee *follows the L-norm* if she acts and votes like kind L. A type-l or L employee *acts and votes with her kind* if she follows the L-norm; analogously for a type-r or -R employee.[1]

At any decision point, let N_L and N_R be the numbers of L-revealed and R-revealed employees. We say that there is an *L-majority* if $N_L > N_R$; an *R-majority* if $N_R > N_L$; and *no majority* if $N_R = N_L$.

Strategies In the third (and final) elimination stage and in choosing a private action, each regular employee always votes and acts according to her own kind. Strong types L and R simply follow a mechanical rule: they act and vote according to their kind in every round, regardless of their perceived type.

It remains to specify strategies for regular employees for decision points up until (but excluding) the third elimination stage. At each decision point, each employee's choice will condition on her type, her kind of perceived type (L-revealed, R-revealed, or unrevealed), and the kinds of perceived types of all current employees. We will distinguish between correctly-perceived revealed employees whose perceived type matches their kind (e.g., a type-r who is R-revealed) and wrongly-perceived revealed employees (e.g., a type-r who is L-revealed). Note that employees are never wrongly-perceived on the equilibrium path.

If there is no majority at the start of an action or elimination stage, then each employee acts or votes with her own kind in that round – with the following exception: in an elimination stage, a revealed but wrongly-perceived employee will vote against nobody.

If there is an L-majority (the case of an R-majority is analogous), then all correctly-perceived type-l's and all unrevealed employees follow the L-norm, while all R-revealed employees (who may be correctly- or wrongly-perceived) act and vote as their type. Finally, consider wrongly-perceived type-r employees. In action stages, if there is a L-majority of at least two ($N_L - N_R \geq 2$), then

[1] Note that as specified, the L-norm does *not* entail voting against unrevealed employees with a slight bias towards R, i.e., perceived type $l/r/R$. This is because an L-norm is meant to reflect "how a regular l-type would act and vote, if unconstrained by norms". Indeed, it would not be in the interest of a regular l-type to eliminate someone who is perceived to be $l/r/R$: at the limit $\gamma \to 0$, the tiny benefit from eliminating that someone (because that someone is very slightly more likely to act as R than L) is outweighed by the enforcement cost ϵ.

they follow the L-norm; with a L-majority of one, they act with their kind. In each elimination stage, wrongly perceived type-r's follow the L-norm – except in the case $N_L = 3, N_R = 2$, where they vote with their kind.

Beliefs We describe how beliefs are updated in action and elimination stages, excluding the final elimination stage. (Given that employees cannot be eliminated by others after the final elimination stage, belief updating at that point has no payoff consequences.)

Consider action stages that start with no majority of either kind. Unrevealed employees who take action L have their perceived type updated by removing r/R from their set of possible types (e.g., from $L/l/r$ to L/l); so these employees become L-revealed. Employees who take action R are updated symmetrically. Revealed employees are not updated regardless of their actions.

Consider action stages that start with an L-majority (the case of an R-majority is analogous). An unrevealed employee who follows the L-norm has her perceived type updated by removing R from her set of possible types (e.g., from $l/r/R$ to l/r); so she remains unrevealed. R-revealed employees do not have their perceived type updated regardless of their action. Finally, an unrevealed or L-revealed employee who violates the L-norm (by taking action R) becomes R-revealed. Specifically: if her perceived type included R, then she becomes perceived to be R; otherwise, she becomes perceived to be r.

Now, turn to elimination stages (except the final elimination stage). Consider elimination stages that start with no majority. Unrevealed employees who vote as kind-L have their perceived type updated by removing r/R from their set of possible types; so these employees become L-revealed. Unrevealed employees who vote as kind-R are updated symmetrically. Revealed employees are not updated regardless of their actions. An unrevealed employee who votes off-path (i.e., neither as kind-L or -R) is perceived to be l; so she becomes L-revealed.[2] A revealed employee's perceived type is never updated regardless of how she votes.

Consider elimination stages that start with an L-majority (the case of an R-majority is analogous). The updating rule here is analogous to that of action stages that start with a majority. R-revealed employees are not updated regardless of their vote. An unrevealed employee who follows the L-norm has her perceived type updated by removing R from her set of possible types; so she remains unrevealed. L-revealed employees who follow the L-norm are not updated. Finally, an unrevealed or L-revealed employee who deviates in any way from following the L-norm becomes R-revealed. Specifically, if her perceived type included R and she deviated by voting like kind R, then she becomes perceived to be R; otherwise, she becomes perceived to be r.

C Equilibrium Existence and Constraints

As Propositions 1a and 1b are independent, we can prove them in either order. It turns out that starting with Proposition 1b is more effective (because we can then reuse its equilibrium existence result to prove Proposition 1a). And so we first turn to Proposition 1b. Our solution concept is perfect Bayesian equilibrium. Let the allowed parameter space (excluding γ) be

$$\Omega = \{(U, V, v, W, w, \epsilon) : U > 0, V > 0, v > 0, W > 0, w > 0, \epsilon > 0\}$$

[2] Notice that the updating rule in this case is asymmetric; we could just have well have specified that an unrevealed employee who votes off-path becomes R-revealed. This assumption is for simplicity of analysis. Specifying a symmetric rule, such as that the employee is randomly revealed to be one kind or another, would not change the rest of the analysis.

and let ω be some generic point in Ω. Let us first reproduce Proposition 1b from the paper.

Proposition 1b *The 'social norm equilibrium' exists for sufficiently large U, small V, large v, small W, and intermediate w, in the following sense:*
There exists a subset S of Ω with non-empty interior such that:

1. *The 'social norm equilibrium' exists for any element in the interior of S*

2. *If $\omega = (U, V, v, W, w, \epsilon) \in S$ then a) $w \geq \epsilon$ and b) if $\omega' = (U', V', v', W', w', \epsilon) \in \Omega$ is such that $U' > U$, $V' < V$, $v' > v$, $W' < W$, $\epsilon < w' < w$ then ω' is also in S*

3. *The 'social norm equilibrium' does not exist for any $\omega \in \Omega$ outside of S*

This statement of the proposition omits the fact that the result focuses on the limiting case $\gamma \to 0$ to simplify the exposition. To make explicit this dependence on γ, we restate Proposition 1b as follows.

Proposition 1b' *There exists a closed subset S of Ω with non-empty interior such that:*

1. *For any element ω in the interior of S, there exists $\underline{\gamma} > 0$ such that the 'social norm equilibrium' exists for all $0 < \gamma < \underline{\gamma}$.*

2. *If $\omega = (U, V, v, W, w, \epsilon) \in S$ then a) $w \geq \epsilon$ and b) if $\omega' = (U', V', v', W', w', \epsilon)$ is such that $U' > U$, $V' < V$, $v' > v$, $W' < W$, $\epsilon < w' < w$ then ω' is also in S.*

3. *For any $\omega \in \Omega$ but outside of S, there exists $\underline{\underline{\gamma}} > 0$ such that the 'social norm equilibrium' does* not *exist for all $0 < \gamma < \underline{\underline{\gamma}}$.*

For the rest of this Appendix, we will focus on Proposition 1b'. The following observation helps to establish the proposition, and also provides some insight into the comparative statics results. We say that an incentive constraint is *relevant* if it deters a one-step deviation from the 'social norm equilibrium'.

Observation 1 *Every relevant incentive constraint takes one of the following forms:*

$$w \geq \epsilon; \tag{1}$$

or

$$U + v \geq \alpha_V V + \alpha_W W + \alpha_w w + \alpha_\epsilon \epsilon + O(\omega; \gamma) \tag{2}$$
$$\textit{for some}$$
$$\alpha_V > 0,\ \alpha_W > 0,\ \alpha_w > 0,$$

where $O(\omega; \gamma)$ is a term that varies across constraints and for which $\lim_{\gamma \to 0} O(\omega; \gamma) = 0$, and where the terms $\alpha_V, \alpha_W, \alpha_w, \alpha_\epsilon$ vary across constraints; or

$$\alpha_U U + \alpha_v v + \alpha_V V + \alpha_W W + \alpha_w w + \alpha_\epsilon \epsilon + O(\omega; \gamma) \geq 0 \tag{3}$$
$$\textit{for some}$$
$$\alpha_U \geq 0,\ \alpha_v \geq 0,\ \alpha_V \geq 0,\ \alpha_W \geq 0,\ \alpha_w > 0,\ \alpha_w + \alpha_\epsilon > 0,\ \lim_{\gamma \to 0} O(\omega; \gamma) = 0,$$
$$\textit{and } \alpha_U + \alpha_v + \alpha_V + \alpha_W > 0;$$

or

$$U + v \geq \alpha'_V V + \alpha'_W W + \alpha'_w w + \alpha'_\epsilon \epsilon + O(\omega;\gamma) \tag{4}$$
$$\textit{for some}$$
$$\alpha'_V \leq \alpha_V,\ \alpha'_W \leq \alpha_W,\ \alpha'_w \leq \alpha_w,\ \textit{with at least one strict inequality,}$$
$$\textit{and } \lim_{\gamma \to 0} O(\omega;\gamma) = 0;$$

where α_V, α_W, α_w are the coefficients for some relevant incentive constraint of the form (2).

There are a finite number of relevant incentive constraints, and there is at least one relevant incentive constraint of each of the forms (1) and (2).

Relevant constraints of the form (1) ensure that regular employees are willing to incur the cost of eliminating opposite-kind employees in the third (final) elimination stage. They imply the comparative-static result that the equilibrium exists only for high-enough w.

Relevant constraints of the form (2) ensure, essentially, that in situations with a majority, the majority (action or voting) norm is followed by unrevealed regular employees of the opposite kind. Notice that the remaining comparative statics results of Proposition 1b' follow from (2): such incentive constraints are satisfied for high U, high v, low V, low W, and low w. To understand these comparative statics, consider the tradeoffs involved for an unrevealed type in violating the majority-kind's norm. Doing so increases the chances of being eliminated, and thus is unattractive if the employee highly values survival (U) and the chance to take a private action (v). On the other hand, survival under a opposite-kind norm entails being hurt, on average, by others' actions (because everyone else conforms to the 'undesirable' norm) and by others' private actions (because an opposite-kind norm means that there are more employees of the opposite kind, on average, than of the same kind). Thus as an employee's dependence on others' actions (W and w) increases, the attractiveness of following the norm and avoiding elimination diminishes. This effect is augmented by the following amplification mechanism: by violating the norm and revealing her kind, an employee may trigger a 'showdown' between the two kinds where everyone reveals their kind with their actions and votes. Such a showdown ends with the focal employee surviving in and only in those states where there is a majority of her own kind – that is, in those states where she benefits on net from the others' actions and private actions. As W and w increase, these potential 'showdown' benefits also increase, and her temptation to violate the norm and trigger such a showdown increases as well.

Relevant constraints of the form (3) and (4) are trivially satisfied as $\gamma \to 0$. They correspond, for instance, to incentive constraints for regular employees of the majority kind not to violate the norm (and thus become perceived to be minority-type employees).

The 'social norm equilibrium' exists if and only if all relevant incentive constraints are satisfied. We will show in Section C.1 that Observation 1 holds. We now prove Proposition 1b' conditional on Observation 1.

Proof of Proposition 1b': [Note that the following argument conditions on Observation 1, and thus is not a self-contained proof of Proposition 1b'.]

We first construct the set S based on (1) and on the set of relevant incentive constraints of the form (2). To start, for any relevant incentive constraint (2) and its associated parameters $\alpha_V, \alpha_W, \alpha_w, \alpha_\epsilon$, define the corresponding "auxiliary" constraint as:

$$U + v \geq \alpha_V V + \alpha_W W + \alpha_w w + \alpha_\epsilon \epsilon, \tag{5}$$

and the corresponding "strict auxiliary" constraint as

$$U + v > \alpha_V V + \alpha_W W + \alpha_w w + \alpha_\epsilon \epsilon. \tag{6}$$

Let S be the intersection of all half-spaces of the auxiliary-constraint form (5) and the half-space corresponding to constraint (1); that is, the set of elements of Ω that satisfy all such auxiliary constraints plus the constraint $w \geq \epsilon$. Correspondingly, S's interior then consists of all $\omega \in \Omega$ that satisfy all the strict auxiliary constraints (6) plus the strict constraint $w > \epsilon$. Note that this interior is nonempty; for instance, pick any $V > 0, v > 0, W > 0, w > 0, \epsilon > 0$ with $w > \epsilon$, and pick sufficiently large U.

It suffices now to show that, given S, all three properties from Proposition 1b' are satisfied. To start, consider Property 1. Consider an element ω in the interior of S. To demonstrate Property 1, it suffices to show that for *any* relevant constraint of one of the forms (1)–(4), ω satisfies that constraint given sufficiently small (but positive) γ. If so, because there are only a finite number of such constraints, we can choose some $\underline{\gamma} > 0$ such that ω will satisfy *all* relevant incentive constraints for all $\gamma < \underline{\gamma}$. That is, Property 1 holds with this choice of $\underline{\gamma}$.

We'll go through each type of constraint in turn. *First*, ω satisfies (1) by definition of S. *Second*, consider any constraint of the form (2). The element ω satisfies the corresponding strict auxiliary constraint (6); therefore, for sufficiently small γ (which implies sufficiently small $O(\omega;\gamma)$), ω must also satisfy (2). *Third*, consider any incentive constraint of the form (3). Suppose (1) holds; then

$$\alpha_U U + \alpha_v v + \alpha_V V + \alpha_W W + \alpha_w w + \alpha_\epsilon \epsilon > 0,$$

so the constraint (3) is satisfied for sufficiently small γ. *Fourth*, consider any relevant constraint of the form (4). We know that, as stated in Observation 2, there is a corresponding relevant constraint of the form (2) which – from above – is satisfied by ω for sufficiently small γ. Further, again for sufficently small γ, (2) is stricter than (4); thus (4) is satisfied. Combining these four points, Property 1 is established.

Now, consider Property 2. Consider any $\omega \in S$ and any ω' as specified in Property 2. Given that $\omega \in S$, it satisfies all relevant auxiliary constraints. Notice that because $\alpha_V > 0, \alpha_W > 0, \alpha_w > 0$, each relevant auxiliary constraint continues to hold if we increase U, increase v, decrease V, decrease W, or decrease w. Thus $\omega' \in S$ also; it follows that Property 2 holds.

Finally, consider Property 3. Consider any point $\omega \in \Omega$ outside S. By our construction of S, either (i) $w < \epsilon$ or (ii) some auxiliary constraint of the form (5) is violated at ω. If (i) holds, then the 'social norm equilibrium' does not exist under ω for any γ. If (ii) holds, then the corresponding relevant incentive constraint (2) is also violated, and the 'social norm equilibrium' does not exist, for sufficiently small γ. It follows that Property 3 holds for ω with a sufficiently small choice of $\underline{\underline{\gamma}}$.
■

C.1 Incentive Constraints

To verify Observation 1, we derived the full set of relevant incentive constraints, and checked that each constraint took one of the forms (1)–(4). These derivations were performed in a brute-force fashion and do not generate much additional insight. Below, instead of reproducing all our calculations for every relevant incentive constraint, we show the calculations for a representative subset of constraints. (The rest of the calculations are available from the authors.) Each subsection discusses a particular class of possible one-step deviations and provides one or two examples.

IC constraints during the final (third) elimination stage, and private actions

We start by checking incentive-compatibility conditions for regular employees to vote their type in the final elimination stage. Specifically, we claim that incentive-compatility holds here if and only if (1) is satisfied.

Consider a type-r employee n's voting choice vis-à-vis another employee m. This choice impacts n's payoff if and only if (i) n is not eliminated, and (ii) n's vote on m is pivotal. In that case, by eliminating m, n incurs a cost of ϵ and ensures $E\left[J_{b_n=\tau_m}\right] = 0$. If m is R-revealed, then he will choose private action R with probability 1; so, absent elimination, $E\left[J_{b_n=\tau_m}\right] = 1$. Similarly, if m if L-revealed, then absent elimination, $E\left[J_{b_n=\tau_m}\right] = -1$. Thus, voting to eliminate an R-revealed m changes n's expected payoff by $-p(w+\epsilon)$ where p is the probability that n survives the vote and that n's vote to eliminate m is pivotal; voting to eliminate an L-revealed m changes n's expected payoff by $p(w-\epsilon)$. It follows that it is never optimal for n to vote against an R-revealed m; it is optimal for n to vote against an L-revealed m if and only if $w \geq \epsilon$. Finally, note that any vote by n against an unrevealed m is never pivotal. To sum up, it is optimal for n to vote as kind R, as the equilibrium prescribes, if and only if $w \geq \epsilon$; i.e., if (1) holds. An identical argument holds if employee n is type-l.

It is obviously optimal for each employee to act with her kind in the private action round, which is the final decision point in the game.

For the rest of this Appendix, we focus on action and elimination stages up till (and including) the third action stage.

IC constraints for unrevealed employees to follow the norm

Suppose there is a majority: say, $N_R > N_L$. In this case, the equilibrium specifies that all regular unrevealed employees will follow the R-norm. We have to separately consider the incentive constraints for (i) a regular l-type (minority) and (ii) a regular r-type (majority) to do so.

Example (i): if the focal employee is of the minority type, these incentive constraints take the form of (2). Focus on (for example) the case of an unrevealed l-type where, at the start of the second action stage, there is an R-majority with $N_R = 2$ and $N_L = 1$. We study whether it is incentive-compatible for the focal employee to follow the R-norm in the second action stage. In the continuation equilibrium, the focal employee follows the R-norm, never reveals herself and thus is never eliminated.

At this point, note that the continuation equilibrium plays out differently, depending on whether any of the other unrevealed employees (including future replacements drawn from the pool) are strong L-types. For a given decision point, let's label an outcome as *atypical* if at least one employee who is present in at least one round of the continuation equilibrium is a strong type; otherwise, we label the outcome as *typical*. We call the first type of outcome 'atypical' because such outcomes are rare – strong-type employees are rarely drawn (when γ is small). Also, note that in typical circumstances, unrevealed employees (being regular) never violate the norm and never subsequently get eliminated. In our example: conditional on a typical outcome, the initially-L-revealed employee will be eliminated and replaced in the second elimination stage, everyone will follow the R-norm in the third action stage, and the initially-R-revealed employees are eliminated with probability 1/4 in the third elimination stage. (Why 1/4? This outcome is realized if and only if both of the non-focal unrevealed employees, including the new replacement, are type-l – which occurs with probability 1/4. In this case, they and the focal employee all vote to eliminate the initially-R-revealed employees.) Thus, conditional on a typical outcome, the expected payoff to the focal employee is (after some computation)

$$U + v - V - \frac{3w}{2} - 4W - \frac{3\epsilon}{2}. \tag{7}$$

What about atypical outcomes? Instead of explicitly tracing out the continuation equilibrium, we bound (coarsely) the focal employee's expected payoffs by considering the lowest and highest possible payoffs any employee can obtain in any history:

$$\begin{gathered} U_n \in [U_{\min}, U_{\max}] \quad \text{where} \\ U_{\min} = \min\{0, U - (V+v) - 4(W+w)\} - 12\epsilon \quad \text{and} \\ U_{\max} = U + (V+v) + 4(W+w). \end{gathered}$$

In addition, we can bound the probability of an atypical outcome. Any unrevealed employee is a strong type with probability $\leq \gamma$, and n interacts with at most two unrevealed employees following the decision point; so the probability of an atypical outcome is at most 2γ. Combining these observations, the focal employee's (unconditional) expected payoff U_n differs from her *typical* expected payoff by at most $2\gamma \cdot (U_{\max} - U_{\min})$; and can thus be approximated (using Big-O notation) as

$$U_n = \left(U + v - V - \frac{3w}{2} - 4W - \frac{3\epsilon}{2} \right) + O(\omega, \gamma) \tag{8}$$

for some function $O(\omega, \gamma)$ where $\lim_{\gamma \to 0} \frac{O(\omega,\gamma)}{\gamma}$ is bounded and so $\lim_{\gamma \to 0} O(\omega, \gamma) = 0$.

On the other hand, suppose the focal employee violates the R-norm in the second action stage and thus becomes L-revealed. Conditional on a typical outcome, the game proceeds to the second elimination stage with $N_R = N_L = 2$. The focal employee survives with probability $1/2$, which occurs if the remaining unrevealed employee also becomes L-revealed – in which case the two R-revealed employees are eliminated. The focal employee's (unconditional) expected payoff is thus, via a similar calculation to (8),

$$\frac{1}{2} (U + v + V + 2w + 4W - 2\epsilon) + O(\omega, \gamma). \tag{9}$$

Note that here, as in (8), the term $O(\omega, \gamma)$ – which varies across constraints – accounts for atypical outcomes. Comparing the focal employee's payoff from following versus violating the norm, we see that following the norm is optimal if

$$U + v \geq 3V + 5w + 12W + \epsilon + O(\omega, \gamma).$$

As claimed, the incentive constraint takes the form (2).

Example (ii): If the focal employee is of the majority type, the relevant incentive constraints take the form (3). For example, suppose that at the start of the second action stage, there is an R-majority with $N_R = 2$, $N_L = 1$. Let's focus on an unrevealed type-r employee, and study whether it is incentive-compatible for her to follow the R-norm in the second action stage. If the focal employee follows the R-norm, then (typically) the L-revealed employee is eliminated in the second elimination stage, and nobody is subsequently eliminated (given that there is an absolute majority for types r and R – consisting of the two R-revealed employees plus the focal employee – in the third elimination stage, as well as zero remaining L-revealed employees). Thus the focal employee's expected payoff is

$$U + v + V + 4W + 2w - \epsilon + O(\omega, \gamma).$$

On the other hand, suppose the focal employee violates the R-norm in the second action stage by taking action L, and thus becomes (wrongly) L-revealed. Typically, in the subsequent (second) elimination stage, everyone votes their kind, except the focal employee who votes against nobody; the focal employee will survive with probability 1/2 (which occurs if and only if the previously-unrevealed employee L-revealed herself); in which case nobody else is eliminated. Note that – if she survives – the focal employee's perceived type is not updated; he remains wrongly L-revealed at the end of the second elimination stage. At the start of the third action stage, $N_R = 2$ and $N_L = 3$. All employees act and vote their kind, the focal employee survives and becomes R-revealed, and the two L-revealed employees are eliminated in the third elimination stage. The focal employee's expected payoff is thus

$$\frac{1}{2}\left(U + V + v + 2w - 2\epsilon\right).$$

Combining observations, it follows that the focal employee follows the R-norm if and only if

$$U + V + v + 8W + 2w + O(\omega, \gamma) \geq 0;$$

as claimed, this incentive constraint takes the form (3).

Examples (i) and (ii) dealt with potential norm violations in action stages. We also have to check that enforcing norms (by voting against revealed employees of the minority kind) is incentive-compatible in elimination stages. The logic is similar to the action-stage case above, but there may be multiple possible ways to violate a voting norm, so we have to check each case. We illustrate with an example.

Example (iii): Suppose $N_R = 2$ and $N_L = 1$ at the start of the second elimination stage. Focus on the case of an unrevealed type-l employee and her decision in the second elimination stage. The 'social norm equilibrium' specifies that she should conform by voting like type R; that is, by voting against the L-revealed employee. If she does so, then: conditional on typical outcomes, the L-revealed employee is eliminated and replaced in the second elimination stage. Everybody follows the R-norm in the third action stage; the two R-revealed employees are eliminated with probability 1/4 in the third elimination stage; and everybody acts like their kind in the private action round. The focal employee is never eliminated. The focal employee's expected payoff is thus

$$U + v - V - \frac{3w}{2} - 4W - \frac{3\epsilon}{2} + O(\omega, \gamma).$$

Now, let's enumerate the possible deviations. The focal employee may vote like type L and thus become L-revealed. In that case, noting that the other unrevealed employee will follow the R-norm and vote against the initially-L-revealed employee, the initially-L-revealed employee (and nobody else) will nonetheless be eliminated in the second elimination stage. The third action stage then starts with $N_R = 2$ and $N_L = 1$; the two unrevealed employees follow the norm; and the focal employee survives with probability 1/4 (in which case the two unrevealed employees reveal themselves to be kind-L, and the two R-revealed employees are eliminated in the third elimination stage). Her expected payoff is thus

$$\frac{1}{4}\left(U + v + V - 4W + 2w - \epsilon\right) - \epsilon + O(\omega, \gamma).$$

The focal employee may deviate off-path in a number of ways. She may vote against nobody; she may vote against one or both of the two R-revealed employees, with or without voting against the L-revealed employee. However, each case produces the same distribution of *typical* outcomes – because the focal employee's votes are never pivotal, whereas each deviation causes her to becomes L-revealed. Specifically, the existing L-revealed employee is voted against by the two R-revealed employees and the remaining unrevealed employee, and thus is always voted off; whereas no other employee receives more than one vote. Thus the typical continuation outcome following each deviation is identical to that if the focal employee had instead voted like type L; and the focal employee's expected payoff following each deviation is identical as well. We conclude that following the norm in the second elimination stage in this example is incentive-compatible if and only if

$$U + v \geq \frac{5V}{3} + 4W + \frac{8w}{3} + \frac{\epsilon}{3} + O(\omega, \gamma);$$

thus the relevant incentive constraint takes the form (2).

IC constraints for unrevealed employees in the absence of a norm

In rounds that start with no majority ($N_L = N_R$), the 'social norm equilibrium' specifies that an unrevealed employee will act, and vote, with her kind. We may check that the incentive constraints at such decision points will take the form (3).

Example (iv): Perhaps the leading example is the first action stage. There, the game always starts with $N_L = N_R = 0$: i.e., no majority. Focus on an unrevealed type-l employee. Suppose, as the equilibrium specifies, she chooses action L. We enumerate the possible *typical* outcomes, and calculate the expected payoff under each outcome:

1. With probability 1/16, all four other employees become L-revealed (i.e., choose action L). In that case, the game ends after the first round, and the focal employee's expected payoff is

$$U + V + v + 4W + 4w.$$

2. With probability 4/16, three of the other employees become L-revealed, and the remaining employee becomes R-revealed. In that case, the R-revealed employee is eliminated and replaced in the first elimination stage; the unrevealed replacement follows the L-norm in the second round; and the game ends then. The focal employee's expected payoff is

$$U + V + v + 4W + 3w + O(\omega, \gamma).$$

3. With probability 6/16, two of the other employees become L-revealed, and the remaining two become R-revealed. In that case, the R-revealed employees are eliminated and replaced in the first elimination stage; the unrevealed replacements follow the L-norm in the second round; and the game ends then. The focal employee's expected payoff is

$$U + V + v + 4W + 2w + O(\omega, \gamma).$$

4. With probability 5/16, at least three of the other employees become R-revealed. In that case, the focal employee is eliminated in the first elimination stage, and receives a payoff of zero.

The focal employee's unconditional expected payoff from taking action L is thus the weighted mean of these payoffs.

On the other hand, suppose the focal employee takes action R in the first action stage, and thus is wrongly R-revealed. The possible typical outcomes are:

1. With probability 1/16, all four other employees become R-revealed (i.e., choose action R). In that case, the focal employee will follow the R-norm in the first elimination stage, and the game ends then. The focal employee's expected payoff is
$$U - V + v - 4W - 4w.$$

2. With probability 4/16, three of the other employees become R-revealed, and the remaining employee becomes L-revealed. In that case, the focal employee will follow the R-norm in subsequent rounds. The L-revealed employee is eliminated and replaced in the first elimination stage, and the game ends after the second round. The focal employee's expected payoff is
$$U - V + v - 4W - 3w + O(\omega, \gamma).$$

3. With probability 6/16, two of the other employees become R-revealed, and the two become L-revealed. In that case, the (wrongly R-revealed) focal employee will vote with her kind L in the first elimination stage, and thus become L-revealed; the R-revealed employees are eliminated and replaced in the first elimination stage; the unrevealed replacements follow the L-norm in the second round; and the game ends then. The focal employee's expected payoff is
$$U + V + v + 4W + 2w + O(\omega, \gamma).$$

4. With probability 5/16, at least three of the other employees become L-revealed. In that case, the focal employee (being wrongly R-revealed) is eliminated in the first elimination stage, and receives a payoff of zero.

The focal employee's unconditional expected payoff from taking action L is thus the weighted mean of these payoffs. Comparing her expected payoff from action L versus action R, we find that it is optimal for her to take action L in the first action stage if and only if
$$\frac{5V}{8} + \frac{W}{2} + 2w + O(\omega, \gamma) \geq 0;$$
as claimed, this incentive constraint takes the form (3).

IC constraints for revealed employees of the minority kind

Consider a decision point where there is a "weak" R-majority: $N_R \geq N_L$. Our equilibrium specifies that any L-revealed employee will act / vote with her kind. Note that this employee's perceived type will not be updated regardless of her action or vote.

In the case that the decision point is an action stage, note that the focal employee's choice of L versus R has no effect on her perceived type, and thus does not change the continuation equilibrium. It follows immediately that it is weakly optimal for her to act with her kind (strictly so if there is a positive probability that Part I will end in that round).

In the case that the decision point is an elimination stage and there is a strict R-majority, $N_R > N_L$, we divide the analysis into two subcases. First, suppose $N_R \geq 3$. Then the focal employee will be eliminated with probability one, and will receive zero payoff regardless of her vote; thus it is weakly optimal for her to vote with her kind. Second, suppose $N_R = 2$ and $N_L = 1$, which can only occur at the start of the second (not the first) elimination stage. Then, given that the two R-revealed employees will vote against the focal employee, she will not be eliminated if and only if the two unrevealed employees are strong L-types and thus do not vote against her. Conditioning on this possibility, it is straightforward to calculate that the incentive constraint for her to vote with her kind takes the form (3).

In the case that the decision point is an elimination stage and there is no majority, $N_R = N_L$, we can show, subcase-by-subcase, that the relevant incentive constraints take the form (3) as well.

IC constraints for wrongly-perceived employees of the minority kind (to follow the norm)

Consider a decision point where $N_L > N_R$, so there is an L-majority. If $N_L \geq N_R + 2$, our equilibrium specifies that any wrongly-perceived majority-revealed (L-revealed, type-r) employee will follow the L-norm. Any incentive constraint for this decision point takes the form (2). We illustrate with two examples.

Example (v): Consider the second action stage. Suppose there is an L-majority with $N_L = 2$ and $N_R = 0$. Focus on a wrongly-perceived, L-revealed (i.e., type-r) employee. If she follows the L-norm by taking action L, then in the *typical* continuation equilibrium, she (and everyone else) will also follow the L-norm in the second elimination stage by not voting against anyone else, at which point the game ends. Her expected payoff is thus

$$U - V + v - 4W - w + O(\omega, \gamma).$$

On the other hand, if the focal employee violates the L-norm by taking action R, then in the typical continuation equilibrium, the second elimination stage starts with $N_L = N_R = 1$, and everyone votes as their own kind. The focal employee survives with probability $1/2$, and her expected payoff is

$$\frac{1}{2}\left(U + V + v + \frac{5W}{2} + \frac{9w}{4} - \frac{7\epsilon}{4}\right) + O(\omega, \gamma).$$

Comparing expected payoffs, it is thus optimal for her to follow the norm if

$$U + v \geq 3V + \frac{21W}{2} + \frac{17w}{4} - \frac{7\epsilon}{4} + O(\omega, \gamma);$$

that is, the incentive constraint takes the form (2), as claimed.

Example (vi): Again, consider the second action stage, and suppose there is an absolute L-majority with $N_L = 4$ and $N_R = 0$. Focus on a wrongly-perceived, L-revealed (i.e., type-r) employee. If she follows the L-norm by taking action L, then in the *typical* continuation equilibrium, she (and everyone else) will also follow the L-norm in the second elimination stage by not voting against anyone else, at which point the game ends. Her expected payoff is thus

$$U - V + v - 4W - 3w + O(\omega, \gamma).$$

On the other hand, if she violates the L-norm, then she is R-revealed, and is eliminated with probability one in the second elimination stage; thus receiving zero payoff. Consequently, it is optimal to follow the L-norm if and only if $U + v \geq V + 4W + 3w + O(\omega, \gamma)$, which indeed takes the form (2).

What if $N_L = N_R + 1$ at a decision point? Then the 'social norm equilibrium' specifies that (a) a wrongly-perceived majority-revealed employee will act and vote with her kind, except if (b) $N_L = 2, N_R = 1$ at the start of an elimination stage – in which case a wrongly-perceived majority-revealed employee will vote as the majority kind. If (a), the incentive constraint takes the form (3); if (b), the incentive constraint takes the form (2). We illustrate (a) and (b) with examples.

Example (vii): Consider the second action stage, and suppose that $N_L = 3$ and $N_R = 2$. Focus on a wrongly-perceived, L-revealed (i.e., type-r) employee. If (as the equilibrium specifies) she votes with her kind R, then in the typical continuation equilibrium: she will become R-revealed, and the two remaining L-revealed employees will be eliminated, in the second elimination stage. Everyone follows the R-norm in the third round, and nobody is eliminated. The focal employee's expected payoff is thus

$$U + V + v + 4W + 2w - 2\epsilon + O(\omega, \gamma).$$

The focal employee may deviate from the equilibrium in multiple ways, depending on how she votes. For this example, we consider the case where she enforces the norm by voting against the two R-revealed employees. Then both R-revealed employees are eliminated; in the third action stage, the focal employee (and everyone else) continue to follow the L-norm; in the third elimination stage, she votes against the two L-revealed employees, who are eliminated with probability 1/4. The focal employee's expected payoff is thus

$$U - V + v - 4W - \frac{3w}{2} - \frac{5\epsilon}{2} + O(\omega, \gamma).$$

Comparing expected payoffs, it is thus optimal for her to play the equilibrium if

$$2V + 8W + \frac{7w}{2} + \frac{\epsilon}{2} + O(\omega, \gamma) \geq 0;$$

that is, the incentive constraint takes the form (3), as claimed.

Example (viii): Consider the second elimination stage, and suppose that $N_L = 2$ and $N_R = 1$. Focus on a wrongly-perceived, L-revealed (i.e., type-r) employee. If she follows the L-norm by voting with her kind L, then in the *typical* continuation equilibrium: in the second elimination stage, the sole R-revealed employee will be eliminated. The focal employee will follow the L-norm in the third round, and will be eliminated together with the other L-revealed employee in the third elimination stage with probability 1/8 (i.e., if all three unrevealed employees turn out to be type r.) The focal employee's expected payoff is thus

$$\frac{7}{8}\left(U - V + v - 4W - \frac{10w}{7}\right) - \epsilon + O(\omega, \gamma).$$

The focal employee may deviate from the equilibrium (by violating the L-norm) in multiple ways, depending on how she votes. However, in each case, the typical continuation equilibrium is the same: in the second elimination stage, the R-revealed employee (and nobody else) is eliminated, and the focal employee becomes R-revealed. Consequently, the focal employee survives the third round with probability 1/2. Her expected payoff is thus

$$\frac{1}{2}\left(U + v + V + \frac{9w}{4} + \frac{W}{4}\right) - \frac{3\epsilon}{2} + O(\omega, \gamma).$$

Comparing expected payoffs, it is thus optimal for her to follow the norm if

$$U + v \geq \frac{11V}{3} + 10W + \frac{19w}{3} - \frac{4\epsilon}{3} + O(\omega, \gamma);$$

that is, the incentive constraint takes the form (2), as claimed.

IC constraints for correctly-perceived, revealed employees of the majority kind (to follow the norm)

Consider a decision point where there is an L-majority: $N_L > N_R$. Our equilibrium specifies that any correctly L-revealed employee (i.e., type-l) will act / vote with her kind. In this case, the relevant incentive constraint takes the form (3).

Example (ix): Consider the second action stage, and suppose there is an L-majority with $N_L = 2$ and $N_R = 1$. Focus on a correctly-perceived, L-revealed (i.e., type-l) employee. If she follows the L-norm by acting like kind L, then in the *typical* continuation equilibrium: in the second elimination stage, the sole R-revealed employee will be eliminated. The focal employee will (with everyone else) follow the L-norm in the third round, and will be eliminated together with the other L-revealed employee in the third elimination stage with probability 1/8 (i.e., if all then-unrevealed employees turn out to be type r). The focal employee's expected payoff is thus

$$\frac{7}{8}\left(U + v + V + \frac{10w}{7} + 4W\right) - \epsilon + O(\omega, \gamma).$$

If the focal employee deviates from the equilibrium (by taking action R) in the second action stage, then she becomes (wrongly) R-revealed, and (typically) in the second elimination stage, the remaining L-revealed employee (and nobody else) is eliminated. At the start of the third action stage, $N_R = 2$ and $N_L = 0$; Consequently, the focal employee follows the norm in the third action stage, then survives the third elimination stage with probability 7/8. Her expected payoff is thus

$$\frac{7}{8}\left(U + v - V - \frac{10w}{7} - 4W\right) - \epsilon + O(\omega, \gamma).$$

Comparing expected payoffs, it is thus optimal for her to follow the norm if

$$\frac{7V}{4} + \frac{5w}{2} + 7W + O(\omega, \gamma) \geq 0;$$

that is, the incentive constraint takes the form (3), as claimed.

D Existence of a Social Norm

We finally consider Proposition 1a, which asserts that the 'social norm equilibrium' exists and further specifies two properties of the equilibrium. Let us start by reproducing the statement of Proposition 1a from the paper.

Proposition 1a *The 'social norm equilibrium' exists (in a non-empty part of the parameter space) and has the following properties:*

1. *On the equilibrium path, there exists (with strictly positive probability) a social norm in at least one round of this equilibrium.*

2. *Whenever this social norm takes the form of an L-norm, the organization will have an L-culture, and analogously for an R-norm and an R-culture.*

We have already shown in Proposition 1b that the 'social norm equilibrium' exists for some parameter values, so it remains to be shown that Properties 1 and 2 from Proposition 1a hold. These two properties may be simply restated as follows.

Observation 2 *In the 'social norm equilibrium', at each stage of the second round and the action stage of the third round:*

1. *Suppose that, on or off the equilibrium path, there is an L- or R-majority and a regular-type employee of the opposite kind who is unrevealed. Then this employee will follow the majority norm.*

   Further, this circumstance occurs with positive probability on the equilibrium path.

2. *On the equilibrium path, with probability one, there is a majority (either L- or R-) of at least 3 players.*

Observation 2 follows immediately from the definition of the 'social norm equilibrium'. Part 1 implies Property 1 from Proposition 1a. It describes circumstances under which a social norm exists: after all, regular-type employees who choose to follow the majority norm despite their opposite preferences do so *only because they would subsequently be eliminated otherwise*. Further, such circumstances arise with positive probability.[3] Note that by inspecting the construction of the social norm equilibrium, one may verify that these are the only circumstances under which a social norm arises on the equilibrium path.

Part 2 of Observation 2, in combination with Part 1, implies Property 2 from Proposition 1a. It points out that on the equilibrium path, any majority that emerges after the first elimination stage will necessarily have at least 3 members – and thus corresponds to our definition of (L- or R-) culture in Section 6. This majority will be maintained for the rest of the game. Given that social norms arise on the equilibrium path only under the circumstances described in Part 1, it follows that if there is a social norm, then there must also be a culture with the same direction (either L- or R-) as the norm.

[3] One example of such circumstances is when (i) an L-majority of less than five players emerges in the first round, so that at least one employee is eliminated in the first round, and (ii) one of the random replacements is a regular r-type employee.