How Many Chiefs? The Role of Leadership in Social Dilemmas

James A Best

January 31, 2011

Abstract

In this paper I show that asymmetric information about payoffs can induce cooperative behavior arbitrarily close to first best in a social dilemma with sequential action. The mechanism that induces such outcomes is 'good leadership'. Leaders are more informed agents whose actions influence the actions of less informed agents due to their informational advantage. This influence can cause leaders to internalise the social cost of their actions and choose socially optimal behavior, set good examples, in social dilemmas. This behavior is then imitated by their followers. However, too many leaders in a population crowd out each others' influence and incentive to be good leaders. Therefore, when too large a proportion of the population is informed leaders do not set good examples and outcomes are Pareto inefficient. Finally, I derive conditions under which ex-ante welfare gains accrue from restricting the proportion of informed agents in a population.

JEL Codes: C72, D62, D82, D83.
Keywords: Asymmetric information, cooperation, efficiency, failure of leadership, good leadership, public goods, prisoners' dilemma, social dilemma.

1 Introduction

In social dilemmas, where externalities imply that private optimization leads to inefficient outcomes, leaders are capable of influencing others to act in a socially beneficial manner through the example of their actions . One of the most prevalent cases of this is where we set examples for children, with varying degrees of success, on issues such as honesty and sharing. Also, experienced workers or bosses may work harder in order to ensure other workers contribute more to a project (Hermalin, 1998). Another case is where the very wealthy donate large sums of money to charities to induce others to make contributions (Andreoni, 2005). These attempts at leadership often go awry. For example, a large purchase order for US Steel by prominent financiers on October 24th 1929 was meant to shore up the markets by influencing other investors to act likewise; it was unsuccessful. There are also occasions where no attempt is made to set a good example; not every parent behaves honestly in front of their child, nor does every boss work extra hard and not every millionaire engages in large public acts of philanthropy. It is not clear why leadership may sometimes induce superior social outcomes and why other times it may not.

In this paper I develop a model of good leadership in which a finite number of agents act sequentially in a social dilemma[1]. Agents are leaders in this model if they have information that other agents do not. Therefore, one could read 'leader' in this paper as meaning 'better informed agent', 'more experienced agent' or 'expert'. [2] It is known if an agent has information or not. Under these conditions agents without information learn from the actions of agents with information. In cases of extreme information asymmetries the less informed agents copy more informed agents. This makes the more informed agents de facto leaders. The measure of whether a leader is a good leader is the extent to which that leader behaves in the socially optimal fashion, sets a good example. The success of good leadership is measured by the number of agents it induces to behave cooperatively.[3]

Using this model of leadership I show four main results. The most significant is that asymmetric information allows leadership to induce cooperative behavior in a proportion of the population in social dilemmas. Moreover, as the population becomes arbitrarily large the proportion of cooperative agents tends to one. Second, the ability of a population to solve social dilemmas through good leadership fails if too large a proportion of the population are leaders. 'Too large' is determined by the relative sizes of the population, the private gains of non-cooperative behavior and the social costs

[1]The game is very similar to a prisoners' dilemma except that there are several alternative strictly dominated actions in the action set other than *cooperate* and *defect*. There are no externalities attached to these dominated actions.

[2]This paper does not pretend to present an encompassing model of leadership. There are many individuals who are not conventionally considered leaders that are considered leaders within this model. The converse is also true, many individuals who are conventionally considered leaders may not be considered leaders in the framework of this model.

[3]In this paper '*cooperate*' and 'socially optimal action' are used interchangeably. Likewise, '*defect*' is used to refer to the action with the highest private payoff and the worst social payoff.

of non-cooperative behavior. Third, there is a critical ratio of the private gains to the social costs of non-cooperative behavior that implies, if it is exceeded, that there can be no cooperative behavior for a population of any size. Finally, I show that there are ex-ante welfare gains from restricting information to a small but positive proportion of the population.

To see how we get these results consider a sequential game where there are many actions and acting without information can be risky. Leaders know the social and private benefits to actions but less informed people, followers, do not know the payoffs to particular actions but do know the distribution of payoffs. Followers will copy a leader when the expected payoff from some other action is lower than copying. This influence, of leaders' actions on the actions of followers, causes leaders to partially internalise the social benefit of an action: because the leader is affected by the externalities generated by the actions of followers. Therefore, leaders may choose the socially optimal action to induce a following of agents to do likewise. However, it is particularly difficult to induce a better social outcome through leading by example when there are many other leaders because each leader has fewer followers in expectation. Leaders will then have less incentive to set a good example as if a leader has few followers the impact of their actions on the leaders payoff will be small relative to when he has many followers. If there are too many leaders they will crowd out one another's influence to such an extent that there the incentive to set a good example will be too small and all leaders will set bad examples. I characterise this game as an 'Example Setting Game' as the informed agents have to decide whether to set a good or bad example and it is these choices that drive the outcomes.

The most notable economic model of leadership is Hermalin (1998). Leadership in Hermalin (1998), and Andreoni (2005), is where some leader with information about the value of contributing to a public good acts before followers who do not have this information. The leader uses their own contribution to signal to other agents the true value of contribution to the public good. There is a unique equilibrium where the leader contributes more than the second-best level to the public good while all followers contribute at the second-best level. In equilibrium the uninformed agents know as much as the leader about the state of the world once the leader has acted.

In my model leaders do not choose actions to signal that a particular state of the world is true. They choose actions because they know that uncertainty over payoffs constrains followers to copy them. Hence, leaders are more influential in my model than in a signalling model of leadership as leaders do not reveal the payoffs to all actions in the action set through their action. The greater level of influence of leaders in this model implies the welfare gains are much larger than in signalling models of leadership where the welfare increase comes solely from the increased contribution of the leader.

A disadvantage of this approach relative to the signalling model of leadership is that it requires a more complex relationship between the action set and payoffs. If followers can infer too much information about the payoffs to other actions from the equilibrium

play of leaders then they will not copy and the leaders will not have an incentive to set a good example. Whether this model or the signalling model is appropriate for some application is therefore contingent on the particular information and payoff structure of the problem. My model cannot be used for the particular public goods problem of the type examined in their papers. However, it is applicable to a large variety of social dilemmas that have similar properties to such public goods problems. A further difference between my model and those of Hermalin and Andreoni is that it examines the effect of multiple leaders on the incentives for good leadership and does not take it as given that the first agent to act will always be the agent with the best information.

This paper uses a framework similar to that expounded in the social learning models of Banerjee (1992) or Bikchandani *et al.* (1992). It may not be correct to characterise my model as a social learning model as such as it is known who has information and the informed agents have perfect information. Learning exists in this model as the uninformed learn from the informed. However, the model does not have the informational externalities present which are of concern in conventional social learning literature. [4]

Another way in which this paper differs from the other social learning literature is that there are externalities to the particular actions of agents which allows for a model of good leadership. There are two other papers, (Dasgupta, 1999; Bhalla, 2007), that attempt to examine externalities in social learning models. However, the results in these papers pertain to the effect on social learning of coordination externalities. They use the same information structure found in the social learning literature. They do not deal with actions that are good or bad for society in themselves. Also, they say nothing about how asymmetric information allows for leadership. Moreover, they do not show how such leadership can lead to Pareto improvements on cases of perfect and symmetric information in the context of social dilemmas.

The folk theorem states that both agents playing *cooperate* is a subgame perfect equilibrium of an infinitely repeated prisoner's dilemma. Likewise, all play *cooperate* can be sustained in a sequential prisoner's dilemma if there is a patient, infinite population with perfect information. In finite games, however, these equilibria break down. The mechanism in this paper for generating cooperative behavior with a finite population is dependent on asymmetric information. This has some similarity to Kreps *et al.* (1982) where asymmetric information over an agents' type can yield partial cooperation in a finitely repeated prisoner's dilemma. However, their model, unlike mine, has no learning or leadership. Moreover, it requires a game to be repeated for cooperation to occur.

Finally, this paper implies that it may be desirable to restrict the public availability of information. There is other work that finds gains from restricting information such as Morris and Shin (2002). However, their welfare results are driven by public information leading to inferior decision-making. This is not the driving factor in my paper.

[4]Preliminary work on a model with imperfect precision of information suggests that cooperative equilibria are still sustainable but that sometimes agents will herd on the wrong action: one that is believed to be cooperative but is in fact the analogue of fishing. Moreover, the welfare optimising proportion of informed agents will be greater the higher the imprecision of the information.

2 A Simple Illustration

Take the following example: there is a tribe of islanders just beginning to interact with a globalised world. The children of the tribe will inherit their own portion of beachfront property when they turn twenty-one. When they inherit their land they will decide what to do with it immediately as they are rather impatient. There are three options for their beach; training in traditional fishing; letting a consortium build a large hotel; or building low impact eco-resorts . Letting the consortium build a large hotel will be the most profitable should no one else do so. However, if everybody builds hotels they will all be worse off than if they build eco-resorts because an ugly coast line caused by everyone building big hotels will drive away tourism. Fishing is a very bad option that will eventually leave them destitute due to the fish stock being destroyed by foreign industrial fishing fleets.

Only one of the islanders knows which actions yield which payoffs. However, the other islanders do know that one of the options is likely to be very bad and know that one of the actions may have large social costs; just not which actions correspond to these outcomes. The islanders do not care about the well being of the others and so each would like to choose the most individually profitable action. The Chief's son, who recently finished university, knows which actions yield which outcomes and all the other islanders know it. It so happens that the Chief's son turns twenty-one before the rest of the islanders and will have to make his decision first. He knows the other islanders know that he will not choose the individually worst option. Moreover, because they are so scared of choosing the worst option[5] they will copy him even if they know he has not chosen the privately most profitable action. Consequently, the Chief's son chooses to build the eco-resort rather than the hotel because he doesn't want to suffer the negative externalities generated by all the islanders that will copy him. Hence, the Chief's son sets a good example because he has internalised the social cost of his action and the islanders get the first best outcome.[6]

However, there is an alternative scenario where the son has a sister at university. She turns twenty-one after exactly half of the islanders have turned twenty-one. It so happens that the payoff from building a hotel when only half of the islanders have built a hotel is larger than that of an eco-resort when no islander has built a hotel. The worst that can happen to the sister if she builds the hotel will be if all the islanders after her build a hotel. This is better than building an eco-resort and therefore she will build a hotel. The islanders after her will all copy her because they know that her action will be at least as selfish as her brothers. However, the son realises this and therefore realises that only half of the islanders will copy him if he builds an eco-resort which is not enough to make him do so instead of building a hotel. Consequently, the

[5]Which, from their perspective, could be either of the two options that the Chief's son has not chosen.

[6]The fact that the Chief's son action is the action that will be chosen by all implies that it is rational, unlike in cases lacking a leadership dimension, for the Chief's son to act according to Kant's categorical imperative.

son decides to build a hotel as does everybody else in the village which yields a worse outcome for all. The future bad influence of his sister crowded out his influence and with it his incentive to set a good example. If the Chief had only sent one child to university they would all be better off. When there are too many potential leaders the leaders will not attempt to set good examples.

In this example the villagers would prefer the sister to have not studied at university. This is because when the leader always acts first no welfare benefits accrue from subsequent leaders. They provide no extra information and crowd out the first leader's incentive to set a good example. However, it was not necessarily the case that the Chief's son would turn twenty-one first. Without his action to guide them the islanders would run the risk of choosing fishing which would be disastrous. Thus one advantage of more children going to university is the greater probability that an educated child acts early and stops others from choosing the very bad option of fishing. This advantage may, or may not, outweigh the effect of more information crowding out the incentive for good leadership. Having more informed agents in the population has two welfare effects cutting in opposite directions. The social benefit of more information is that less agents choose very poor actions with low individual payoffs. The social cost is that more agents will behave uncooperatively in expectation.

3 The Example Setting Game

An example setting game is a sequentially played social dilemma with asymmetric information about payoffs. A finite population of n agents are drawn sequentially from an infinite general population N. Each agent acts after being drawn from the general population and before the next agent is drawn. The action of agent i is a_i and is chosen from the action space $A = [0, 1]$. The profile of all actions up to and including a_i is denoted as A_i. Two elements $c, d \in [0, 1]$ are chosen by 'Nature' with uniform probability before the first agent is drawn. c and d are the only elements in the action space with positive instant payoffs. The instant payoff from d is larger than the instant payoff from c. However, there is a negative externality generated by d. After the nth agent has acted all agents receive a negative 'externality payoff' proportional to the number of agents who have chosen d. [7] This gives the payoff to agent i as a function of the action profile of the population, A_n, as:

$$U_i(A_n) = u(a_i) - \sum_{j=1}^{n} \varepsilon_j \tag{1}$$

where the instant payoff from action a_i is given by:

[7] The pertinent equilibrium results also hold for finite action sets of three or more actions and for richer distributions of instant payoffs and externalities. This is examined further in the penultimate section 'Alternative Payoff and Externality Distributions'.

$$u(a_i) = \begin{cases} z & \text{if } a_i = c \\ v & \text{if } a_i = d \\ 0 & \text{if } a_i \notin \{c, d\} \end{cases} \tag{2}$$

and :

$$\varepsilon_j = \begin{cases} \varepsilon > 0 & \text{if } a_j = d \\ 0 & \text{if } a_j \neq d \end{cases} \tag{3}$$

and the following conditions on the payoffs hold:

$$v - z > \varepsilon > 0 \tag{4}$$

$$v - (z + \varepsilon) = \Delta > 0 \tag{5}$$

$$L = n \cdot \varepsilon - v + z \tag{6}$$

ε is the negative payoff to other agents from playing d and will be referred to as an externality. 'Own-action payoff' of action a_i is used to refer to the instant payoff less any negative end of game payoff to i from action a_i. This will be equal to an action's instant payoff for all actions except d. d has an own-action payoff of $v-\varepsilon$.[8] Condition (5) implies that this is larger than the own-action payoff of choosing c. The difference between the two own-action payoffs of c and d, Δ, is referred to as the 'defection incentive'. Hence, c and d are akin to *cooperate* and *defect* in a prisoners' dilemma. [9] Throughout this paper c and d shall be referred to as '*cooperate*' and '*defect*'. The difference in the instant payoffs to c and d must be less than the product of the externality and the population size for this to be a social dilemma. This is the case where condition (6) holds. L denotes the magnitude of the Pareto loss from the population choosing an all *defect* rather than an all *cooperate* action profile. L can be thought of as a measure of the size of the social dilemma that the population faces. Consequently, *cooperate* is always Pareto superior to all *defect*. This game is akin to an n person sequential prisoners' dilemma.

Information in the game is asymmetric. There are two types of agents: Leaders and Followers. Proportion α of the general population are Leaders and proportion $1 - \alpha$ are Followers where $0 \leq \alpha \leq 1$. Leaders observe Nature's move at the beginning of the game and Followers do not. Therefore, only leaders have direct knowledge of the own-action payoffs to all actions. The type of agents[10] and their actions are public

[8]It makes for easier exposition to have the externality from choosing d also affect the agent who chooses d. It makes no difference to the results as d is still individually optimal.

[9]To see this consider the game where $n = 2$. In this case $a_1 = a_2 = c$ is Pareto superior to $a_1 = a_2 = d$ but d is the dominant strategy under perfect information.

[10]It is not necessary for agent types to be known for a cooperative equilibrium to be sustained but it does make for a simpler proof.

knowledge subsequent to having chosen an action. Note, that Followers only observe the number that Leaders choose and do not observe whether this number corresponds to c, d or neither. Thus the information set h_i of any agent i must contain the action profile A_{i-1} and the types of all agents up to and including i. All agents know the value of α and the payoffs attached to an agent correctly choosing c or d.

4 Equilibrium Play

Agents' actions can have an effect on the play of subsequent agents. This implies the choice of action affects the expectation of the end of game externality payoff. Therefore, agents' actions maximize the expectation of the own-action payoff less its 'externality impact'. The expected externality impact of action a_i is the expectation of the externalities generated in equilibrium by subsequent agents conditioned on h_i and a_i.

In this paper expectations and actions will be determined within a Bayesian Perfect equilibrium. Thus, 'equilibrium' in this paper refers to Bayesian Perfect equilibrium. Followers infer information about the payoffs of different actions from equilibrium play. Followers update their beliefs about payoffs after each Leader's action. The actions of Followers, however, do not affect beliefs in equilibrium as they have no private information. Therefore, Followers' equilibrium expectations of the payoffs of actions will be determined only by the actions of Leaders.

The only uncertainty for Leaders is the externality impact of their actions. In equilibrium this is a function of the types of all agents subsequent to a Leader. The expected externality impact of the action of a Leader is determined, in equilibrium, by the expected size of that Leader's Following. The Following of Leader i is all the Followers acting after i and before k. k is the first Leader to act subsequent to Leader i. Those Followers acting subsequent to Leader i and before Leader k will, on the equilibrium path, have the same beliefs about the payoffs of actions. This follows from the fact that Followers' actions reveal no information about the payoffs of actions. The expected size of some agent i's Following will be important in later analysis and is given in Lemma 1 below.[11]

Lemma 1 *The expected size of agent i's Following is:*

$$F_i^e = \frac{(1-\alpha) - (1-\alpha)^{n+1-i}}{\alpha}. \tag{7}$$

Equilibrium play is examined here in three parts. The first part looks at play in the case where all agents are Leaders and the implication of the 'Pandora Effect' for pure

[11] See the Appendix for proof.

strategy equilibrium. It is shown that on the equilibrium path all agents play d if $\alpha = 1$. Also, that where it is a pure strategy equilibrium for an agent to play d then all subsequent agents play d for any value of α. The second part is to establish the conditions under which Leaders will definitely play d, or definitely not play d, for a given set of parameters. It is found that Leaders will defect within a finite distance from the end of the game in all equilibria. They will never defect before this point in any pure strategy equilibria. Finally, a pure strategy equilibrium will be shown to exist in which leaders cooperate up to a point and defect thereafter.

4.1 The Pandora Effect

If an agent's action has no effect on subsequent agents' actions then d is optimal. For d maximises the own-action payoff and all actions have equal externality impacts. In the degenerate case of the game, where $\alpha = 1$, all agents play d. This follows from backwards induction. The expected difference in the externality impacts of any two actions is 0 for the nth agent. Therefore, $a_n = d$ independently of history as it maximises own-action payoff. Consequently, the expected difference in the externality impacts of any two actions is 0 for agent $n-1$. Therefore, $a_{n-1} = d$ independently of history also. The same argument then applies for $n-2$ and so on back to the first agent. This outcome is Pareto inferior to all agents choosing c. It is also Pareto inferior to all agents' choosing $x \notin \{c, d\}$ if $n \cdot \varepsilon > v$. In the latter case a world of complete ignorance is Pareto superior to one of perfect information.

Play in the degenerate game, where $\alpha = 1$, implies the following proposition holds for $\alpha \in [0, 1]$:

Proposition 1 (The Pandora Effect) *If $a_i = d$ for Leader i in any pure strategy equilibrium then all subsequent agents will play $a_j = a_i = d \; \forall j > i$.*

Proof. If Leader i plays d in a pure strategy equilibrium it will be believed by all subsequent agents on the equilibrium path that $a_i = d$. This belief is common knowledge. There can be no equilibrium strategy of subsequent Leaders that can change the course of play. This would imply changing the beliefs of Followers from the true belief that $a_i = d$ to a false belief that $a_i \neq d$, and we can't have false beliefs in equilibrium. Therefore, all agents subsequent to i will play according to the degenerate game with $\alpha = 1$ and a population size of $n - i$. Hence, each agent j subsequent to i plays d: $a_j = a_i = d \; \forall j > i$. □

The revelation of information through equilibrium play is irreversible on the equilibrium path. Hence, all subsequent agents will defect after a Leader has defected on the equilibrium path of a pure strategy equilibrium. This is because they all know *how* to defect: the knowledge of how to do bad cannot be put back in the box. This will yield a total game externality of $\varepsilon \cdot (n + 1 - i)$ where i is the first Leader to play d.

4.2 Defection

Leader i will play *defect* when own-action payoff net of expected externality impact is greater for d than c or some $a_i \notin \{c, d\}$. I will show that this is the case for i in any equilibrium where the following inequality holds: i's expected following is smaller than the ratio of the defection incentive[12] to the magnitude of the externality. I define the 'defection condition', D, as holding for agent i where this inequality weakly holds. D strictly holds when this inequality strictly holds. More formally:

D holds if and only if $F_i^e \leq \dfrac{\Delta}{\varepsilon}$.

D strictly holds if and only if $F_i^e < \dfrac{\Delta}{\varepsilon}$.[13]

One property of this condition is that if D holds for agent i then it also holds for agent $i+1$. If D doesn't hold for agent i then it doesn't hold for agent $i-1$. Therefore, if the number of agents for which D holds in the population is Ψ there is a critical agent $\theta = n - \Psi$ for which the following Lemma holds:

Lemma 2 *D holds for an agent i if and only if $i > \theta = n - \Psi$.*

The following proposition uses condition D to give sufficient conditions for defection by Leaders and Followers.

Proposition 2 *On the equilibrium path in all equilibria each of the following are sufficient conditions to imply $a_j = d$:*

1. *j is a Leader for whom D strictly holds.*
2. *j is subsequent to some i who is a Leader for whom D strictly holds.*

Proof. In the following proof of Proposition 2 any statement about the actions of Leaders is conditioned on D strictly holding for that leader.

Consider an equilibrium where Leaders' actions can only affect the actions of their own Following. Let j be the first Leader subsequent to Leader i. The Pandora effect

[12] Recall that the defection incentive is the difference in the own-action payoffs of *cooperate* and *defect*. Which is the difference in payoffs of the two actions should these actions not affect the actions of any other agent.

[13] Recall that F_i^e, the expected Following of i, is the expected number of Followers acting after i but before the next leader. Therefore, a more intuitive way of arranging this inequality may be $\varepsilon \cdot F_i^e < \Delta$. The expected externality generated by i's Following, if they all choose d, is less than the difference in the own-action payoffs of c and d. Which implies that the Leader i's expected payoff from d is greater than c if condition D strictly holds and only i's Following is influenced by i's action.

implies all agents subsequent to j will play $a_j = a_i = d$ if $a_i = d$ is a pure strategy for i. In such an equilibrium the maximum difference in the expected payoffs of *defect* and *cooperate* for i is if all i's Following copy and play a_i. This would yield an expected difference in utility of:

$$E[U_i(d) - U_i(c)] = \Delta - \varepsilon \cdot F_i^e \tag{8}$$

This is strictly positive where D strictly holds. Therefore if D strictly holds for i in such an equilibrium then i would prefer d to c and, *a fortiori*, all other actions. In equilibria where j's strategy is the same but the Following of i have a strategy other than play a_i then the value of equation (8) will be even larger and the reason to play *defect* will be greater. This result can now be used to derive Proposition 2.

If the last agent is a Leader then $a_n = d$ is the optimal action for n in all equilibria for any history of actions. Therefore, from the above argument, if $n-1$ is a Leader and D holds then $a_{n-1} = d$ in all equilibria. From the Pandora effect this implies $a_n = a_{n-1} = d$ in all equilibria where $n-1$ is a Leader for which D strictly holds. By backward induction this holds for $n-2$ if D strictly holds and so on till the first Leader for which D strictly holds. Thus, if D strictly holds for Leader i on the equilibrium path of any equilibrium then $a_i = d$. The Pandora effect then implies that $a_j = a_i = d$ for all $j > i$. This concludes the proof of Proposition 2. □

Proposition 3 *There is no pure strategy equilibrium where some leader, for whom D does not hold, plays d on the equilibrium path.*

Proof. If it is a pure strategy for the first leader, i, to play d then the Pandora Effect implies that all subsequent agents play d on the equilibrium path. If i chooses an action other than d then i's following still play a_i as they are unaware that they are off the equilibrium path. In such a case the minimum that i gains from playing c is if only i's Following copy a_i and all subsequent agents play d. In this case the minimum expected gain from cooperating instead of defecting is $\varepsilon \cdot F_i^e - \Delta$. If condition D doesn't hold for i this gain is larger than zero. Therefore it cannot be an equilibrium for the first leader to play d if D doesn't hold. If it is a pure strategy for the second leader, j, to play d then j has a minimum expected gain of $\varepsilon.F_i^e - \Delta$ from playing c. This follows as Leader j's following will not play d because they will not know which option is d on the equilibrium path as the first leader can't have played d either. Proposition 3 then follows from forward induction. □

A corollary of Propositions 2 and 3 is:

Corollary 1 *The number of agents who can possibly defect in a Bayesian pure strategy equilibrium is bounded above by Ψ: the number of agents for which D holds. Therefore,*

all agents acting up to and including agent $\theta = n - \Psi$ will not play defect in any pure strategy equilibrium.

4.3 An Equilibrium

There is an equilibrium where Leaders up to and including the critical agent θ cooperate and then defect thereafter. In this equilibrium each Leader is copied by their Following. This is formalised in the proposition below.

Proposition 4 *There is a pure strategy equilibrium consisting of the following strategies and out of equilibrium beliefs.*

Leader i's strategy is:

$$a_i = \begin{cases} c & \text{if } i \leq \theta \\ d & \text{if } i > \theta. \end{cases} \tag{9}$$

Follower j's strategy is to play $a_j = a_i$ if they are in the Following of Leader i and to play $a_j = 0$ if there is no Leader $i < j$.

Off the equilibrium path beliefs are:

1. *All Followers in Leader i's Following believe that $a_i = d$ regardless of the actions of previous Leaders.*
2. *All Followers who deviate from the equilibrium path are believed to be uninformed.*

Proof. The off-path beliefs of Followers about Leaders implies that it is rational for Followers to copy the last Leader off the equilibrium path. Therefore, in this equilibrium, Leader i's action will affect the actions of i's Following alone. Therefore, where D holds it will be optimal to defect and where D does not hold it will be optimal to cooperate. Thus the decision rules in Proposition 4 maximize Leaders expected utility. The equilibrium play by the Leaders implies that the actions of Followers subsequent to the first Leader maximizes expected own-action payoff as agents acting before the first Leader subsequent to θ cannot choose d. Those acting after the first Leader subsequent to θ play d from the Pandora Effect; which is what adopting the above strategy entails.

Off the equilibrium path beliefs imply that any deviation by a Follower will not change the number of other agents playing d. Therefore the expected externality impact is at the same as when they play the equilibrium strategy. Therefore copying the previous Leader's action maximizes their expected utility as it maximizes their own-action payoff. Before the first leader acts then follower's expected own-action payoffs will

be 0 from any action. Deviation from the equilibrium path will not change the expected externalities. Therefore, choosing action 0 is weakly preferred to any other action by followers acting prior to the first leader. This concludes the proof of Proposition 4. □

I have shown in Proposition 2 that all equilibria will imply the same ex-ante play after agent θ as in the equilibrium above. However, I have not shown that all equilibria imply all leaders cooperate before and up to agent θ. There are two possible sets of equilibria that may exist which would need to be ruled out to show this.

The first set are possible pure strategy equilibria in which Leader $i \leq \theta$ plays $a_i \notin \{c, d\}$. This may be sustained if a Leader prior to θ is punished by subsequent agents for playing c. There may be off the equilibrium path beliefs that would make such a punishment strategy incentive compatible. Such equilibria would be Pareto inferior to the equilibrium above as expected externalities are the same but average own-action payoffs are lower. The second set are possible equilibria in which leaders up to and including θ play mixed strategies. However, neither set of equilibria are ruled out in this paper.

5 The Conditions for Cooperation

In this section I look at what conditions on the parameters imply that a given game is expected to yield cooperative behavior. Moreover, I look at how much cooperative behavior can exist in such a game. I find two main results for the equilibrium defined above. The first is that there exists a critical proportion of informed agents above which there will be no cooperative agents in a population of any size. The second is that below this critical level the proportion of agents that will choose to cooperate tends to one as the population becomes large.

Theorem 1 *In the equilibrium defined in Proposition 4 there will be no cooperative agents in a finite population of any size if the proportion of informed agents is at or above a critical level $\bar{\alpha}$, where:*

$$\bar{\alpha} = \frac{\varepsilon}{\varepsilon + \Delta} \tag{10}$$

Proof. If D holds for the first agent then D holds for all subsequent agents. D holds for the first agent if:

$$F_1^e = \frac{(1-\alpha) - (1-\alpha)^n}{\alpha} < \frac{\Delta}{\varepsilon} \tag{11}$$

F_1^e is monotonic and increasing in n which gives the unclosed upper bound of F_1^e as $\frac{1-\alpha}{\alpha}$. Therefore, there is no finite population size for which D will hold for any agent when the following equation holds.

$$\frac{1-\alpha}{\alpha} \leq \frac{\Delta}{\varepsilon} \tag{12}$$

Which always holds where $\alpha \geq \bar{\alpha}$. □

Theorem 2 *If $0 < \alpha < \bar{\alpha}$ then as $n \to \infty$ the expected proportion of the population that cooperates tends to one.*

Proof. There are two sets of agents that will not choose *cooperate*. First are those agents acting before the first Leader who I refer to as 'the Ignorant'. Second, agents for which D holds and are either a Leader or acting subsequent to a Leader for which D holds, 'the Defectors'.

The expected size of the first group is bounded above by the expected size of the Following of the first agent in a with a population size of $n+1$. This is increasing in n and bounded above by $\frac{1-\alpha}{\alpha}$ which is finite for $1 > \alpha > 0$.

The size of the second group is bounded above by Ψ from Corollary 1. $\alpha < \bar{\alpha}$ implies that $\frac{1-\alpha}{\alpha} > \frac{\Delta}{\varepsilon}$. Therefore, there exists some finite integer X such that $F_1^e \leq \frac{\Delta}{\varepsilon}$ for $n = X$ and $F_1^e > \frac{\Delta}{\varepsilon}$ for $n = X+1$. Ψ cannot be greater than X; thus the total number of agents that would choose to play d in equilibrium is bounded above by X for any n.

Hence, the expected proportion of agents playing c in equilibrium is bounded below by:

$$\frac{n - (X + \frac{1-\alpha}{\alpha})}{n} \tag{13}$$

The upper bounds for the Ignorant and the Defectors are invariant with n and hence equation (13) tends to one as $n \to \infty$. □

6 Welfare and the Proportion of Informed Agents

I show in this section that it can be better, in the ex-ante welfare sense, to have neither perfect information nor complete ignorance. There are sufficient conditions that imply

expected welfare is greater when the proportion of informed agents, α, has an interior value rather than a corner value.

Lemma 3 below gives the average expected welfare.[14]

Lemma 3 *The expected average welfare as a function of α is:*

$$W_{\alpha}^{e} = P\frac{z}{n}[n - \Psi - I_{\theta-1}^{e} + F_{\theta}^{e}] + (\frac{v}{n} - \varepsilon)[\Psi - F_{\theta}^{e}] \tag{14}$$

$P = 1 - (1 - \alpha)^{\theta}$ is the probability of some agent being a Leader who is not a Defector. $I_{\theta-1}^{e}$ is the expected number of ignorant agents conditional on some agent being a leader who is not a Defector.

The expected average welfare in the corner cases of $\alpha = 0$ and $\alpha = 1$, W_0^e and W_1^e, are:

$$W_0^e = 0 \tag{15}$$

$$W_1^e = v - \varepsilon \cdot n \tag{16}$$

The relative sizes of v, ε and n determine whether expected welfare is higher in the case of perfect information or of complete ignorance. Expected welfare is higher in a game of complete ignorance than a game of perfect information where $v - \varepsilon \cdot n < 0$. Expected welfare is higher in a game of perfect information than a game of complete ignorance where $v - \varepsilon \cdot n > 0$. In the first case, the loss due to externalities dominates the private gain from agents being able to choose d. In the second case, the loss of positive private payoffs from having no information dominates the cost of externalities generated by agents choosing d. These two cases are called 'externality dominated games' and 'information dominated games' respectively. There are sufficient conditions for an interior α to yield higher ex-ante expected welfare in both externality dominated games and information dominated games.

In an externality dominated game the cost to society of some agent defecting is very high. This cost will be incurred with positive probability if some portion of the general population is informed as some agent will defect with positive probability. However, a small number of Leaders may still induce a better outcome through their good example, c, than no Leaders at all. For this to be the case, the expected cost of externalities generated by defectors must be dominated by the welfare gains from the majority being able to get a positive instant payoff of z. Proposition 5 gives sufficient conditions for this to be the case.

[14]See the Appendix for proof.

Proposition 5 [15] *It is the case that there is some* $0 < \alpha < 1$ *for which* $W^e_\alpha > W_0 > e > W^e_1$ *when:*

$$v - \varepsilon \cdot n < 0 \tag{17}$$

and condition (18) holds:

$$z > [(n-1)\varepsilon - \Delta] \frac{\underline{X} - F^e_{n-\underline{X}}}{P[n - \underline{X} - I^e_{n-\underline{X}-1} + F^e_{n-\underline{X}}] + \underline{X} - F_{n-\underline{X}}} \tag{18}$$

There always exists a large finite n^* *such that condition (18) holds when condition (19) holds:*

$$n > n^* \text{ and } z > \varepsilon. \tag{19}$$

The right hand side of condition (18) in Proposition 5 is always finite for small enough $\alpha > 0$. This implies that if the payoff from *cooperate* is sufficiently large, relative to the extent that the externality dominates the defection incentive, then we prefer to have some informed agents in the population. The private payoff z can be attained by Leaders without causing negative externalities. Therefore, z can be seen as the pure benefit of information. It is that component which can be realised in an individuals payoff without causing lower payoffs for others. While $L = (n-1)\varepsilon - \Delta$, the size of the social dilemma, can be interpreted as the cost of information. If the pure benefit of information is large enough, that is if z is large relative to L, then having some information in the population is always preferred to having none. Condition (19) implies that this is always the case in a large population when the externality inflicted on society by one person playing *defect* is less than the value of all agents being able to choose *cooperate*.

Now I consider the case of an information dominated game. The relevant point of comparison for an information dominated game is the case of perfect information where the whole population is informed, $\alpha = 1$. In this case a reduction in the proportion of informed agents implies a positive expected number of ignorant agents who will not choose d. This, relative to the case of a fully informed population, yields an average loss of $v - \varepsilon \cdot n > 0$ for each ignorant agent. Take the case where the reduction is small enough that condition D still holds for all agents. In this case expected average welfare is strictly lower. However, if D no longer holds for some agents then the the loss from the ignorant has some compensation. The loss if offset by gains from agents playing *cooperate* where they would have previously played *defect*. Each agent who plays c instead of d implies a gain of $(n-1)\varepsilon + \Delta = L > 0$.

The overall effect of reducing α to a level that induces some cooperation is ambiguous. I would like a general solution for when this will be welfare improving or not; such

[15] See the Appendix for proof.

a solution has been elusive. I only look at the specific case where the proportion of informed agents is such that D holds with equality for the second agent. In this case D does not hold for the first agent who, if a Leader, will play *cooperate*. The proportion of informed agents that implies D holds with equality for the second agent is defined as α_{n-1}. I am able to derive the following proposition for the case where $\alpha = \alpha_{n-1}$.[16]

Proposition 6 *Where $v - \varepsilon \cdot n > 0$ there is some $0 < \alpha < 1$ for which $W_\alpha > W_1 > W_0$ if the following inequality holds:*

$$L - 1 - \alpha_{n-1} z > 0 \tag{20}$$

where α_{n-1} is decreasing in L.

Hence, the larger the size of the social dilemma, L, the larger the value of z that implies imperfect information is preferred to perfect information. As noted earlier z can be seen as a measure of the value of information net of externalities. Therefore, the larger this value of information the larger the social dilemma needs to be to imply that we would rather not have all members of the population informed in an information dominated game.

7 Alternative Action Spaces and Payoff Structures

The action set, instant payoff distribution and the nature of the externalities in the model outlined above make for easy exposition of the equilibrium results but are restrictive. In this section I show that the equilibrium results hold for a much less restrictive set of assumptions. I do not relax any other aspect of the game: the number of agents, the proportion of Leaders in the general population, the order of action and the information sets of agents remain the same.

In the model above I have the action space A, the instant payoff function $u(.)$ and the externality function $\varepsilon(.)$. Here, I replace them with A', $u'(.)$ and $\varepsilon'(.)$ respectively. The action space A' is any measurable subset of the real line; this allows for finite or continuous action spaces. In the action space there are still two singletons, c and d, analogous to *cooperate* and *defect*. I now allow the instant payoffs given by $u'(.)$ and the externalities given by $\varepsilon'(.)$ to be such that actions in A' other than c or d can have non-zero payoffs and/or have externalities. Also, I now allow the externalities given by $\varepsilon'(.)$ for c and d to be positive or negative.[17] Leaders still know the exact mapping from the action space to the payoff and externality space while Followers do not. Followers still know what the instant payoffs and externalities are even though they do not know

[16] See the Appendix for proof.

[17] Note that in the set up that a positive value given by $\varepsilon'(.)$ implies a negative externality and a negative value given by $\varepsilon'(.)$ implies a positive externality.

the mapping. It may be the case that the instant payoff or externality of an action is correlated, in a non-trivial sense, with the instant payoffs or externalities of other actions. This allows Followers to learn more from the actions of Leaders than they were previously able[18].

In the following proposition I outline sufficient conditions on A', $u'(.)$ and $\varepsilon'(.)$ under which the equilibrium results of the paper still hold.[19]

Proposition 7 *Propositions 1 to 4 and Theorems 1 and 2 hold for any action set A', instant payoff function $u'(.)$ and externality function $\varepsilon'(.)$ if all the following conditions hold:*

$$E_j[u'(a_j = a) - \varepsilon'(a_j = a)|h_j, a_i = c] < u'(c) - \varepsilon'(c) \ \forall a \in A'_{-c}, \ \forall i \leq \theta \ and \ \forall j \in F_i, \tag{21}$$

$$u'(d) - \varepsilon'(d) > u'(a) - \varepsilon'(a) \ \forall a \in A', \tag{22}$$

$$u'(c) - \varepsilon'(c) \geq u'(a) - \varepsilon'(a) \ \forall a \in A'_{-d}, \tag{23}$$

$$\varepsilon'(c) < \varepsilon'(a) \ \forall \ a \in A'. \tag{24}$$

Condition (21) states that the expected own-action payoff of an action other than c is strictly less than the own-action payoff of c for any Follower who has played subsequent to Leaders who have only played c. For condition (21) to hold it must be the case that the action space admits a minimum of three possible actions. Moreover, there must be at least one action that yields a worse payoff than *cooperate*. Furthermore, it also implies that the payoff function cannot be invertible from equilibrium play and Leaders' strategy functions.

Conditions (22) and (23) imply that *defect* and *cooperate* still have the first and second highest own-action payoffs. The first of these conditions is a matter of definition. If d does not yield the highest own-action payoff then agents, in a perfect information case, will want to choose some action other than d. We would then wish to call this other action d instead. Conditions (22) and (23) also imply that the difference in own-action payoffs for *cooperate* and *defect* is positive. This implies there is an incentive to

[18]For example, with the action set $A' = \{1, 2, 3\}$ a Follower may believe that 1 is d with probability p if $3 = c$ and believe it is d with probability $q \neq p$ if $2 = c$. The Follower would be able to infer which element was c from equilibrium play and therefore update his expectations for the payoffs to playing $a = 1$ accordingly.

[19]See the Appendix for proof.

play defect, the defection incentive, which is equivalent to Δ in the standard model. Consequently, the relevant consideration for Leaders deciding whether to play *cooperate* or *defect*, conditional on being copied by just their Following, is still the Defection Condition. Condition (24) states that *cooperate* generates the most positive externality of all actions and is therefore the socially optimal action [20].

Finally, the combination of conditions (23) and (24) warrants further discussion. In conjunction they imply that the socially best action is necessarily the individually second best. This does not allow for cases where the actions that have the second to nth best own-action payoffs are socially worse than c. It seems reasonable to expect a spectrum of actions socially worse than c that yield higher own-action payoffs. If *both* (23) and (24) must hold for the equilibrium results to go through then the results of this paper will only apply to a small subset of social dilemmas. However, preliminary investigations show that where either (23) or (24) do not hold the equilibrium results are qualitatively similar to the results of the standard model. This is not shown here formally but it is easy to intuitively see why the results will be much the same.

In the standard model the Leader's change from c to d when the impact on externalities of d becomes small relative to the defection incentive, Δ. Now, consider the case where there is a set of less cooperative actions $\{d_1, d_2, ..., d_n\}$ such that the own-action payoffs are monotonically increasing in the index on d but all have a higher own-action payoff than c. Also, the externality generated by these actions is worse than the externality generated by c and monotonically increasing[21] in the index on d. In this new case there is a defection incentive Δ_1 between c and d_1, Δ_2 between d_1 and d_2, and so on till d_n. There is also a larger externality from moving to a higher indexed d which translates to a higher externality impact from moving to a higher indexed d. Leaders will move from one action to the next when the difference in the expected externality impacts of those actions becomes small relative to the defection incentive between those actions.

This is exactly like the standard model except instead of Leaders' strategies being to jump, after some cut-off point, straight from c to d they gradually progress from c to d_1 and upwards through the index to d_n. Hence, agents will cooperate till a point and then cooperate a bit less; and then a bit less; and so on. This is a slow deterioration in cooperative behavior rather than an instant collapse. However, it is still the case that completely cooperative behavior will exist up to a finite distance from the end and that the population will tend towards being completely cooperative for arbitrarily large

[20] In order for this to be a social dilemma it also needs to be the case that the following condition holds:

$$u'(d) - u'(c) < n(\varepsilon'(d) - \varepsilon'(c)). \tag{25}$$

[21] We do not need to consider the possibility of a d_j and d_i where $j > i$ and the own-action payoff of d_j is greater than d_i while the externality is smaller. This is because in such a case there would never be any reason to play d_i as d_j dominates it in both dimensions: instant payoff and externality impact. So, for the purposes of considering equilibrium play we can throw d_i out of the set of less cooperative actions.

populations.

8 Summary and Implications

In social dilemmas with risky action sets the opportunity for good leadership exists if information is restricted to a small proportion of the population. This can induce results, in expectation, that are arbitrarily close to Pareto efficient. Under perfect information we get an inefficient outcome. This cooperative behavior happens within the context of a finite game which distinguishes it from the cooperative results possible in infinitely repeated games. Moreover, if too many agents have access to information the incentive for leaders to choose socially optimal actions is crowded out and the population fails to improve on the second-best outcome; too many leaders causes bad leadership.

Ideally, a policy maker would like the informed agents to all act first as this maximises the number of agents choosing the socially optimal action. These considerations may have been influencing British policy makers during the Swine Flu epidemic of 2008. Doctors and nurses, perceived leaders, were given financial and social incentives to act first in taking the swine flu vaccine. There was an incentive not to take the vaccine as it may have had risky side effects. Moreover, the marginal benefit of taking a vaccine is decreasing in the proportion of the population taking the vaccine as 'herd immunity' develops[22]. The general population was uncertain as to whether it was best to take the risky vaccine or not. If medical staff were able to delay acting they may try and persuade everyone else to take the vaccine so they might free-ride on the herd immunity of the population. Arguably, it was for this reason that incentives were given to ensure that medical staff would act first.

In this model the order of action is exogenous. I take it that there is a stochastic element to the order in which any person is presented with a decision and that there are significant costs to them delaying their decision. However, when this is not the case this model still helps us get some insight into agents' preferences over when to act given the order of everyone else's action. When the dilemma is due to negative externalities then everybody wants to act last. Leaders can choose the privately optimal action and not be followed. Followers have the chance of copying a late leader and getting the higher private payoff. When the dilemma is due to positive externalities it is different because the positive externalities only accrue if a leader shows followers what to do. Consequently, leaders weakly prefer either to be the very first or the very last. If there is a leader acting early in the game then all other leaders would prefer to act last. Then they would accrue the gains from the contributions of others while choosing to play *defect* themselves. However, if all the leaders were acting late then each leader would prefer to act at the beginning of the game rather than the end. This would initiate

[22]Herd immunity occurs when an epidemic can't spread because a sufficiently large proportion of the population is immune.

cooperative behavior and allow the leader to benefit from the positive externalities generated by subsequent followers. Finally, uninformed agents will always try to wait till after some leader has acted as they get a low expected payoff when acting without the guidance of a leader.

One prediction of this model is that we ought to see cooperative behavior occurring most frequently amongst groups where there are a few clearly well informed people and a large number of ignorant people who fear the consequences of independent action.[23] Therefore, it may be desirable to restrict information to a small number of agents and introduce greater uncertainty about the payoffs to particular actions. Information could be restricted by making it costly to attain; making education expensive or making important information difficult to decipher, somewhat like academia. Another way of restricting information is to make it costly to experiment with different actions and behaviors. For example, one could announce that a large punishment exists for some randomly selected action and only allow a few people to know which action incurs such a punishment. Agents will be afraid to experiment with actions because of the risk of choosing the action with the large punishment.

This latter mechanism is reminiscent of the way that conformity or leadership is induced in brutal police states or in the military. However, while such policies may improve social outcomes in my model this is only on the assumptions of perfect knowledge amongst leaders and homogeneous payoffs for all agents. This constrains leaders' choices to be both accurate and, when setting a good example, in line with the interests of everyone else. Whether this is an accurate characterisation of the leaders is very difficult for followers to verify. Therefore, followers ought to be highly dubious of any attempt by leaders to stop them from acquiring information on the grounds put forward in this model. There is no guarantee that the leaders want the information to be restricted for welfare optimizing reasons or for private gain at the cost of the followers. Without such a guarantee it is not clear that followers can accept such a reason as the argument could then be used again and again at the convenience of those agents with a monopoly on information. Even if the payoffs had been homogenous in the first instance it is difficult to see how they would remain homogenous in the long run. [24]

[23]For example, it might imply that large groups of children with a few adults ought to demonstrate more cooperative behavior than large groups of adults with a few children. I think that the relative ignorance of children makes them good subjects for examining this model as the strategies that adults use to deal with social circumstances are often adopted by children and these strategies have a great deal of persistence into later life. Hence, adults when being observed by children may think it wise to ensure that these children learn strategies that are beneficial to the adults. Practising honesty, fairness and concern for the elderly may be wise if this is a strategy that will be imitated by those younger than yourself when you are elderly.

[24]Arguments are often had over the extent to which information available to departments of government is available to the general public. A common defence is that this information is kept secret for the good of the people. A common complaint against this argument is that the people do not know that the government has their best interests at heart or are competent at dealing with this information. The argument for secrecy here is different to my own but the argument against secrecy is similar to the one outlined above.

The welfare results of this model, as noted earlier, are based on leaders having information that is perfectly precise. If the information becomes less precise the less we are able to restrict information in order to yield welfare gains. Therefore this model has implications for the extent to which we should have division of information. There is a trade off between how many agents can be informed and how well informed any one individual might be. Simplistically, we can imagine a situation where there are many fields of knowledge and we can choose to have everyone know a little about each field or have a few experts in each field. Without externalities to actions it is conceivable that there will be cases where it's better to have more people with less precise information in all fields than a few people in each field with precise information. In my model however we have an additional argument for having a smaller number of agents with precise information in each field. Consequently, one might see the results of this paper as an argument for specialisation of knowledge.

A final implication of this model is that leaders may want to hide their actions or the fact that they are a leader. In the second case it turns out that good example setting equilibria still exist where followers believe that someone seeming to be a follower is a fake when they do not follow a leader on the equilibrium path. Consequently, they follow the fake, this induces the leader to play the same equilibrium even if it's not known that they are a leader. However, the first case would imply that with negative externalities all leaders would pay a premium to hide their action. For example, a manager may go to great lengths to shirk in a fashion that cannot be observed by his employees.

Appendix

PROOF of Lemma 1:

The expected number of agents in a leader's following is given by:

$$F_i^e = (1-\alpha)^{x+1}(x+1) + \alpha \sum_{j=1}^{x} j(1-\alpha)^j \tag{26}$$

where

$$n - (i+1) = x$$

The summation term on the right hand side of (26) gives:

$$\begin{aligned}
\sum_{j=1}^{x} j(1-\alpha)^j &= (1-\alpha) + 2(1-\alpha)^2 + ... + x(1-\alpha)^x \\
&= \sum_{j=0}^{x-1}(1-\alpha)^{x-j} + \sum_{j=0}^{x-2}(1-\alpha)^{x-j} + ... + \sum_{j=0}^{x-k}(1-\alpha)^{x-j} + ... + \sum_{j=0}^{0}(1-\alpha)^{x-j} \\
&= \sum_{k=1}^{x}\sum_{j=0}^{x-k}(1-\alpha)^{x-j} \\
&= \sum_{k=1}^{x}\left[\frac{(1-\alpha)^k - (1-\alpha)^{x+1}}{\alpha}\right] \\
&= \frac{(1-\alpha)^1 - (1-\alpha)^{x+1}}{\alpha} + \frac{(1-\alpha)^2 - (1-\alpha)^{x+1}}{\alpha} + ... + \frac{(1-\alpha)^x - (1-\alpha)^{x+1}}{\alpha} \\
&= \frac{\sum_{j=1}^{x}(1-\alpha)^j - x(1-\alpha)^{x+1}}{\alpha}
\end{aligned}$$

Substituting this back into (26) gives:

$$
\begin{aligned}
F_i^e &= (1-\alpha)^{x+1}(x+1) + \alpha\left[\frac{\sum_{j=1}^{x}(1-\alpha)^j - x(1-\alpha)^{x+1}}{\alpha}\right] \\
&= (1-\alpha)^{x+1}(x+1) + \sum_{j=1}^{x}(1-\alpha)^j - x(1-\alpha)^{x+1}\alpha \\
&= (1-\alpha)^{x+1} + \sum_{j=1}^{x}(1-\alpha)^j \\
&= \sum_{j=1}^{x+1}(1-\alpha)^j = \sum_{j=1}^{n-i}(1-\alpha)^j
\end{aligned}
$$

PROOF of Lemma 3:

There are two mutually exclusive and exhaustive sets of possible outcomes. The first set is where some agent $i \leq \theta$ is a Leader. The second set is where all $i \leq \theta$ are Followers. The first set happens with probability $P = 1-(1-\alpha)^\theta$ and the second set with probability $1-P = (1-\alpha)^\theta$.

In the first set of outcomes the expected number of agents acting before a Leader or the expected number of Ignorant agents, $I^e_{\theta-1}$, is given by:

$$
I^e_{\theta-1} = F^e = \sum_{i=1}^{\theta-1}(1-\alpha)^i = \frac{(1-\alpha)-(1-\alpha)^\theta}{\alpha} \tag{27}
$$

The expected number of agents acting subsequent to θ who choose c and not d due to acting before the first uncooperative Leader is the Following of θ will be[25]:

$$
F^e_\theta = \sum_{i=1}^{\Psi}(1-\alpha)^i = \frac{(1-\alpha)-(1-\alpha)^{\Psi+1}}{\alpha} \tag{28}
$$

The expectation of the total externality generated is the product of the externality and the number of uncooperative agents less the expected number of uncooperative agents acting before the first leader. This is the same in both sets of worlds and is given by:

$$
\varepsilon[\Psi - F^e_\theta] \tag{29}
$$

Therefore the expected average welfare conditional on the first set of outcomes is:

[25]Of course θ may not be a Leader in which case it is the hypothetical Following of θ.

$$\frac{z}{n}[n-\Psi-I^e_{\theta-1}+F^e_\theta]+(\frac{v}{n}-\varepsilon)[\Psi-F^e_\theta] \tag{30}$$

Conditional on the second set of outcomes expected average welfare is:

$$(\frac{v}{n}-\varepsilon)[\Psi-F^e_\theta] \tag{31}$$

This gives an unconditional expected average welfare of:

$$W^e=P\frac{z}{n}[n-\Psi-I^e_{\theta-1}+F^e_\theta]+(\frac{v}{n}-\varepsilon)[\Psi-F^e_\theta] \tag{32}$$

PROOF of Proposition 5:

Define $\underline{X}$ as the number of agents in the population for which D holds when $\alpha=0$. In which case D doesn't hold for agent $n-\underline{X}$ and does hold for agent $n+1-\underline{X}$. In the case of $\alpha=0$ the expected following of agent i is $n-i$. As D holds for agent $n+1-\underline{X}$ then $(F^e_{n+1-\underline{X}}|\alpha=0)=(\underline{X}-1)\varepsilon\leq\Delta$. As D doesn't hold for agent $n-\underline{X}$ then $(F^e_{n-\underline{X}}|\alpha=0)=\underline{X}\varepsilon>\Delta$. Therefore, there must be some $\alpha=\underline{\alpha}>0$ for which:

$$\varepsilon(F^e_{n-\underline{X}}|\alpha=\underline{\alpha})=\varepsilon\frac{(1-\underline{\alpha})-(1-\underline{\alpha})^{\underline{X}+1}}{\underline{\alpha}}=\Delta \tag{33}$$

This implies that if $\alpha\in[0,\underline{\alpha})$ if $\underline{X}\in[\frac{\Delta}{\varepsilon},\frac{\Delta+\varepsilon}{\varepsilon}]$. Hence, expected welfare is greater for $\alpha\in(0,\underline{\alpha})$ than $\alpha=0$ when the following inequality holds.

$$(W^e_\alpha|\alpha\in(0,\underline{\alpha})=P\frac{z}{n}[n-\underline{X}-I^e_{n-\underline{X}-1}+F^e_{n-\underline{X}}]+(\frac{v}{n}-\varepsilon)[\underline{X}-F^e_{n-\underline{X}}]>0 \tag{34}$$

This rearranges to give:

$$\frac{z}{n}>(\varepsilon-\frac{v}{n})\frac{[\underline{X}-F^e_{n-\underline{X}}]}{P[n-\underline{X}-I^e_{n-\underline{X}-1}+F^e_{n-\underline{X}}]} \tag{35}$$

Let:

$$K=\frac{[\underline{X}-F^e_{n-\underline{X}}]}{P[n-\underline{X}-I^e_{n-\underline{X}-1}+F^e_{n-\underline{X}}]} \tag{36}$$

Inserting this into equation (35) and using the fact that $v=z+\Delta+\varepsilon$ we get:

$$\frac{z}{n} > (\varepsilon - \frac{z+\Delta+\varepsilon}{n})K \tag{37}$$

Which rearranges to give:

$$z > ((n-1)\varepsilon - \Delta)\frac{K}{1+K} \tag{38}$$

which yields condition (18) in Proposition 5.

The following of agent $n - \underline{X}$ is invariant in the size of n for $n > \underline{X}$. The expected number of Ignorant agents is bounded above by the finite value $\frac{1-\alpha}{\alpha}$. Also, as $n \to \infty$ then $P \to 1$. Hence, as n becomes arbitrarily large condition (18) becomes:

$$z > \varepsilon \tag{39}$$

This yields condition (19) in Proposition 5. □

PROOF of Proposition 6:

The welfare of a game with $\alpha = \alpha_{n-1}$ is:

$$W^e_{\alpha_{n-1}} = \alpha_{n-1}\frac{z}{n}[1 + F^e_1] + (\frac{v}{n} - \varepsilon)(n - 1 - F^e_1) \tag{40}$$

Welfare for α_{n-1} is greater than for perfect information when $W^e_{\alpha_{n-1}} - W^e_1 > 0$. This gives us:

$$\begin{aligned}
W^e_{\alpha_{n-1}} - W^e_1 = \alpha_{n-1}\frac{z}{n}[1 + F^e_1] + (\frac{v}{n} - \varepsilon)(n - 1 - F^e_1) - (\frac{v}{n} - \varepsilon)n &> 0 \quad (41)\\
\alpha_{n-1}\frac{z}{n}[1 + F^e_1] - (\frac{v}{n} - \varepsilon)(1 + F^e_1) &> 0 \quad (42)\\
\alpha_{n-1}z[1 + F^e_1] - (v - n\varepsilon)[1 + F^e_1] &> 0 \quad (43)\\
\alpha_{n-1}z - (z + \varepsilon + \Delta - n\varepsilon) &> 0 \quad (44)\\
\alpha_{n-1}z - (z + \Delta - (n-1)\varepsilon) &> 0 \quad (45)\\
(n-1)\varepsilon - \Delta - 1 - \alpha_{n-1}z = L - 1 - \alpha_{n-1}z &> 0 \quad (46)
\end{aligned}$$

This is the inequality in equation (20).

The relationship between α_{n-1} and L: if $\alpha = \alpha_{n-1}$ then $F^e_2 = \sum_{i=1}^{n-1}(1-\alpha_{n-1})^i = \frac{\Delta}{\varepsilon}$. If n is held constant $\frac{\varepsilon}{\Delta}$ increases then L increases and $\frac{\Delta}{\varepsilon}$ decreases. This implies that

α_{n-1} increases to mantain the above equality. Likewise, if $\frac{\varepsilon}{\Delta}$ is held constant and n increases then α_{n-1} increases to mantain the above equality and L increases. Hence, if L increases α_{n-1} increases. □

PROOF of Proposition 7:

The argument in the proof of The Pandora Effect relies only on d yielding the highest own-action payoff of all actions in A. Hence, Proposition 1 holds if condition (22) holds.

The argument in the proof of Proposition 2 relies on the relationship between the own-action payoffs of c and d defined in condition (5). Condition (5) states only that the own-action payoff from d is greater than the own-action payoff of c: this is equivalent to conditions (22) and (23). It also relies on the expected following size of a Leader; this is independent of A', $u'(.)$ and $\varepsilon'(.)$. Finally, it relies on the Pandora Effect which has been shown to hold under condition (22). Therefore, Proposition 1 holds when conditions (22) and (23) hold.

The arguments in the proofs of Propositions 3 and 4 rely on Leaders' Followings copying their Leaders. This occurs in equilibrium where copying yields the highest expected own-action payoff for Followers. This is necessarily the case for Followers in the Following of a leader playing defect. It is also the case for Followers in the Following of a Leader playing cooperate as a pure strategy if condition (21) holds. Conditions (22), (23) and (24) imply it cannot be optimal for Leaders to play any action other than c or d. Therefore, it must be optimal, conditional on being copied by their Following, for Leaders for whom D does not hold to play c. Hence, Propositions 3 and 4 hold if conditions (21), (22), (23) and (24) hold.

Finally, the arguments for Theorems 1 and 2 depend only on Propositions 1 to 4. Hence, Theorems 1 and 2 hold where conditions (21), (22), (23) and (24) hold. □

References

Andreoni, J. (2005) "Leadership Giving in Charitable Fund-Raising." Journal of Public Economic Theory 8(1): 1-22

Bhalla, M. (2007) "Endogenous Timing, Payoff Externalities and Informational Cascades." Working paper.

Banerjee, A. V. (1992)."A Simple Model of Herd Behavior." The Quarterly Journal of Economics 107(3): 797-817.

Bikhchandani, S., D. Hirshleifer, I. Welch (1992). "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." Journal of Political Economy 100(5): 992.

Dasgupta, A. (1999) "Social Learning with Payoff Complementarities." Working paper.

Hermalin, B. (1998). "Toward an economic theory of leadership: Leading-by-example." American Economic Review 88, 1188-1206.

Kreps, D., P. Milgrom, J. Roberts, R. Wilson (1982) "Rational cooperation in the finitely repeated social dilemma ." Journal of Economic Theory 27(2): 245-252.

Morris, S. and H. S. Shin (2002). "Social Value of Public Information." The American Economic Review 92(5): 1521-1534.