




ARTICLE

Individual versus group morality: the role of information

Sambit Mohanty¹ and Jaideep Roy² 

¹Indian Institute of Technology, Mumbai, India and ²University of Bath, Bath, UK
Corresponding author: Jaideep Roy; Email: jr2014@bath.ac.uk

(Received 17 November 2023; revised 01 April 2024; accepted 30 April 2024)

Abstract

A rise in the number of moral individuals in a group can hurt the morality of the group's collective action. In this paper, we characterize strategic environments and models of morality where this is true solely because, after all, individual morals are private information.

Keywords: Morality; individuals; groups; information; 'Samaritan's Curse'

1. Introduction

Aggregation of individual morality to discern whether a group of individuals is together moral or not is a conceptually difficult task. In a recent work, Basu (2022) has demonstrated that when morality concerns caring for the powerless, and group morality is judged through the consequentialist prism based on how group activity affects the welfare of the powerless, a philosophically exciting phenomenon arises, which he calls the 'Samaritan's Curse'. A rise in the number of moral individuals in a group, possibly due to a sermon by 'the good Samaritan', can worsen the welfare of the powerless who depend solely on the actions taken by the group. This paper exploits this perverse phenomenon by characterizing a fairly large class of games to highlight the critical role that information plays on the welfare of the powerless when the profile of individual morality is not common knowledge amongst the powerful.

A traditional debate between universality and subjectivity of individual morals exists in various works on moral philosophy (see for example Hume (1740) and Kant (1764)). The purpose of this paper is not to understand whether morality is a rational consequence (often yielding a universal code) or is limited to personal sentiments (that yields subjective ethics). Our analysis remains agnostic to these important philosophical questions. Our premise is rather practical in nature where we assume that individual morals may or may not be publicly observable. As

individuals, we may have a particular moral preference, but this may not necessarily be fully known to others with whom we interact.

There is a very large literature on morality in mainstream economics founded in the ideas of other-regarding preferences or in procedural reciprocity.¹ Basu's fundamental departure is to consider the idea of power seriously, a marker that differentiates people between powerful actors whose actions collectively affect payoffs of all individuals, and 'bystanders' whose actions produce no real impact even on their own payoffs; they are the powerless. The celebrated Dictator Game (see Engel (2011) for a meta-analysis and further references) has features of Basu's ideas of divisions based on power where a single decision maker, the dictator, must split an amount between herself and a bystander who, unlike in the Ultimatum Game, has no power to reject an unfair offer and thereby punish the dictator. Basu's environment is a full generalization of this idea where (a) the powerful are not alone but are interlocked with other equally powerful players with conflict of interest and (b) the game between the powerful group and the bystander is not necessarily zero-sum. These two aspects make Basu's world greatly complex when compared with the standard Dictator Game, which after all, is a simple decision problem once the preferences of the dictator are specified, though useful to uncover innate traits about fairness.

Studying societies thus modelled is critical to our understanding of the uncompromising challenges we face today, including international politics over poverty, conflicts between technology, environment and climate where future generations are the powerless, or debates on curbing uncontrolled spread of artificial intelligence to protect the powerless ordinary citizens from the unknowns of 'artificially intelligent' societies. Awareness campaigns and 'moral-policing' around such issues are on the rise, and perhaps even with the right intentions. Basu's observation underscores the harm that these activities can inflict, making our study a natural first step towards evaluating the efficacy of 'public sermons', particularly when it is realistically impossible to judge moral convictions of powerful entities with conflict of interests.

We consider a population of powerful agents (or players in short) with privately observed morals that either make them enjoy an additional utility if the welfare of the powerless rises (viz. the 'morals'), or that keep them unconcerned about the powerless (viz. the 'amorals'). They are randomly matched into pairs to play a 2×2 strategic-form game.² The outcomes achieved by these pairs affect the expected (or total) welfare of the powerless bystanders and our objective is to study how this welfare moves as the proportion of moral players rises.

Basu chooses a particular 2×2 example to demonstrate that in a world of complete information, while the bystander is better off when both players turn moral, she is worst off with only one moral player. We start by showing that such an

¹Some important contributions in this area are Kahneman et al. (1986), Camerer and Thaler (1995), Schwartz (1996), Fehr and Schmidt (1999), Henrich et al. (2001), Fehr and Gächter (2002), Fehr and Schmidt (2006), Basu (2010) and Cooper and Kagel (2016).

²Pairwise random matching is a popular framework used in evolutionary theory. It captures the idea that as individuals living in large societies, we often encounter strangers with whom we must interact in strategic environments (or games).

immiserization of bystander welfare continues to hold in that example when we move to a population game. Since our objective is to understand the critical role of incomplete information in the present discourse, our first goal is to characterize some minimal restrictions on the universe of 2×2 games and the associated payoff matrix of the bystander such that complete information never generates the Samaritan's Curse.³ As the restricted class of games in general allows for multiple equilibria under complete information irrespective of the morality-types of the players (a feature not present in Basu's carefully constructed example with a unique pure strategy equilibrium), we stick to pure strategy equilibria and employ risk-dominance (à la Harsanyi and Selten 1988) for equilibrium selection wherever needed.

As desired, we first establish that under such minimal restrictions if the players learn each other's morality types prior to choosing their actions, a rise in the fraction of moral individuals (weakly) increases the expected welfare of the powerless no matter how the (weakly increasing) 'warm glow' that a moral player experiences is modelled. An important game theoretical point to note here is that morality affects individual payoffs of the powerful players in a non-affine manner, thereby leading to non-trivial consequences on the strategic aspects of the transformed game. Characterizing a large class of games to fit our purpose is therefore tricky, but we find that very reasonable restrictions give us complete information worlds that escape the Samaritan's Curse. Such a universe of games serves as a baseline for our comparison with the case when morality remains private information.

We then show that with incomplete information, the welfare of the bystander continues to increase in the proportion of moral individuals in the population if, for example, his payoff is incorporated linearly in the utility of powerful moral players. However, if (and only if) the moral players are sufficiently *euphoric* about bystander welfare – the additional utility they enjoy from a rise in the bystander's welfare is sufficiently convex so that the utility return to the powerful from a rise in the welfare of the bystander rises at an increasing rate – there exists a population of moral and amoral individuals such that the expected welfare of the bystander decreases as the proportion of moral agents increases. Thus, we provide a partial characterization of the pair of payoff matrices (jointly for the powerful players and the bystander) and the warm-glow function (that incorporates the utility to the powerful from the welfare of the bystander) such that the Samaritan's Curse appears exclusively because morality is not publicly observable.

The intuition behind this is that sufficiently euphoric moral players become so overly concerned about the bystander's payoff that they choose to play a bystander-benevolent action despite the associative risk of mis-coordination with a partner who might be amoral, a risk that even the bystander himself isn't willing to take. Of course, when the proportion of moral players is sufficiently large there is hardly any risk while when this proportion is sufficiently low, there is hardly anyone to take this distortionary risk. This is precisely why the Samaritan's curse is obtained for intermediately moral populations of powerful players.

The rest of the paper is structured as follows. In section 2 we revisit Basu before we describe the framework formally in section 3 where we present a preliminary

³Clearly, Basu's 2×2 example falls outside our restrictions.

result for the complete information world. Section 4 establishes our main results and the paper concludes in section 5. Some proofs are provided in the Appendix.

2. Revisiting Basu

Basu (2022) asks the following question: is it always true that a group with more moral individuals acts in a way that results in a more ‘moral’ outcome than a group which has relatively less moral individuals? To address this, Basu studies a non-cooperative society where a group of powerful individuals take actions in a simultaneous-move strategic environment and their actions together affect the welfare of an individual or a group of powerless bystanders. In that setting, Basu calls an individual moral if she or he enjoys an extra ‘utility’ that increases in the welfare of the bystanders. Through some numerical examples it is then shown that the answer to this question is surprisingly in the negative, leading to what Basu calls the ‘Samaritan’s Curse’, where sermons from a good Samaritan to convert the powerful into moral individuals who are then concerned about the welfare of the powerless, hurts the welfare of the powerless.

The general intuition behind Basu’s finding is rooted in the various nuances of (Nash) equilibrium analysis that have largely been overlooked outside the economics literature in general and in other studies of group morality in particular. It is well known in game theory that if payoffs of all players are transformed affinely, that is by adding and/or by multiplying a fixed constant term that remains common in all outcomes of the game (but not necessarily common across players), there is no effect on individual incentives. Hence, such payoff transformations have no impact on the strategic features of the game, maintaining the same equilibrium set as before. The trouble with morality as modelled in Basu is that when ordinary (*viz.* amoral) players become moral, their payoffs are not necessarily affinely transformed because the bystanders’ welfare consequences can vary in an unorderedly fashion with the game outcomes. As a consequence, the equilibrium set in the transformed game can change, thereby yielding surprising results such as the Samaritan’s Curse.

While Basu’s exposition is strongest when each powerful player has three strategies, as in that case, the welfare of the bystander diminishes with the number of moral individuals, a glimpse of the curse persists in his example where each has just two strategies instead. Since two-strategy two-player games (*viz.* often referred in the game theory parlance as ‘ 2×2 games’) are the most fundamental models in game theory, our focus is restricted to these games.

We begin with Basu’s 2×2 example that yields (partial) possibility of the Samaritan’s Curse in the following sense. Two powerful players 1 and 2 play the left-panel simultaneous-move game depicted in Table 1 except if they are moral, in which case the payoff of the moral player gets raised (additively) by the earnings of the bystander given in the right panel. Using a linearly additive transformation in the above example, Basu demonstrates that in a full-information world where the morality-types of players 1 and 2 are common knowledge, the payoff of the bystander when 1 is amoral and 2 is moral is lower than when both are amoral, but highest when both players are moral. To see this, first note that when neither player

Table 1. The base game *A* and the bystander's payoff matrix *B* in Basu's example

		Player 2	
		X	Y
Player 1	X	100, 101	100, 100
	Y	101, 100	101, 101
		Bystander	
		X	Y
X		2	8
Y		0	4

is moral, the unique Nash equilibrium is where each plays the strategy *Y* that yields a payoff of 4 to the bystander. When both become moral, and therefore the bystander's welfare is added to their original payoffs to determine the transformed utilities of the two players, they play effectively the following game:

where now the new Nash equilibrium is where both players play *X* that yields a

		Player 2	
		X	Y
Player 1	X	102, 103	108, 108
	Y	101, 100	105, 105
		Player 2	
		X	Y
Player 1	X	102, 101	108, 100
	Y	101, 100	105, 101

payoff of only 2 to the bystander. Thus even under complete information, a rise in the number of moral players in a group can be harmful for the bystander.

Basu's 2×2 example can be easily extended to the realm of population games with the following features. Powerful players are randomly matched into pairs from a large population where a fraction μ are moral. In Basu's example then, if each player observes whether the other player is moral or not once they are matched into their respective pairs, the expected 'equilibrium' payoff of the bystander equals

$$EU_B = 8\mu^2 + 2\mu(1 - \mu) + 4(1 - \mu)\mu + 4(1 - \mu)^2$$

as a fraction μ^2 of the population will form pairs between two moral players that yields a payoff of 8 to the bystander, a fraction $\mu(1 - \mu)$ of the population will form pairs where only Player 1 is moral and each player will play *X* that yields a payoff of 2 to the bystander, and so on. The expected 'equilibrium' payoff EU_B of the

bystander is convex in μ reaching its minimum at $\mu = 1/6$. Thus, starting with a $\mu < 1/6$, a rise in μ reduces the bystander's welfare.

Since our objective is to demonstrate that a perverse impact of morality such as the Samaritan's Curse can just be driven by incomplete information, we shall work with general 2×2 games but restrict payoff matrices in such a way that the above two observations will not hold true in a world of complete information. In such a world, our goal is to find conditions where a rise in morality can harm the bystander only when there is lack of information about individual morality between the powerful players at the time these players choose their strategies.

3. Framework and Preliminaries

We consider a society with a continuum of powerful players who are randomly matched in pairs to take part in a 2×2 simultaneous-move (non-cooperative) game. The row and column players in a match are called 1 and 2 respectively and the pure strategy set for each player is $S_1 = S_2 = \{X, Y\}$. At each pure strategy profile $s \in S = S_1 \times S_2$, the material utility to player i is $u_i(s)$. The society further consists of a powerless bystander whose welfare $v(s)$ depends on the 'outcome' s in the game the powerful play.

A fraction μ of the population of powerful players are moral (they care about the welfare of the bystander), while the rest are amoral (they care only about their own payoff from the game). We define a *warm-glow function* $m : \mathbb{R} \rightarrow \mathbb{R}$ to capture the additional payoff $m(v)$ a moral player gets as a function of the bystander's payoff v . Thus, if the bystander's welfare at strategy profile s is $v(s)$, then a moral player obtains a utility of $m(v(s))$, that is, morality is an internalization of the positive externality enjoyed by the bystander. By definition of warm-glow, m is non-decreasing in v , that is, as the welfare of the bystander increases, so does the utility of a moral player. We say that m incorporates *euphoria* if m is convex. In other words, a powerful moral player's preferences exhibit euphoria if the additional utility they enjoy from a rise in the bystander's welfare increases at a rate that itself rises with the rise of the bystander welfare.

Warm glow, or its absence, determines the *psychological payoff* π_i of a player i as follows: at any strategy profile $s \in S_1 \times S_2$,

$$\pi_i(s) = \begin{cases} u_i(s), & \text{if } i \text{ is amoral,} \\ u_i(s) + m(v(s)), & \text{if } i \text{ is moral.} \end{cases} \quad (1)$$

Table 2 depicts the two functions $u(\cdot)$ and $v(\cdot)$ (referred to as matrices A and B respectively), where $a_i, b_i, c_i, d_i, \alpha, \beta, \gamma$ and δ are arbitrary real numbers:

Two informational structures: For the purpose of this paper, we look at two informational structures, one where formed pairs of players learn each other's psychological payoffs $\pi_1(\cdot)$ and $\pi_2(\cdot)$ before playing the game, and the other where $\pi_i(\cdot)$ remains private information to player i . We find a stark difference between these two environments when it comes to the bystander's welfare. In particular, we show that for any non-decreasing warm glow function, a rise in μ necessarily increases the welfare of the bystander under complete information while under incomplete information, this is not true.

Table 2. The base game A (left) and the bystander's payoff matrix B (right)

		Player 2	
		X	Y
Player 1	X	a_1, a_2	b_1, b_2
	Y	c_1, c_2	d_1, d_2
		Bystander	
		X	Y
X		α	β
Y		γ	δ

As discussed extensively in section 2, the addition of $m(v(\cdot))$ does not necessarily lead to an affine transformation of u_i and therefore can change the strategic features of the game in disorderly ways. To obtain clear theoretical predictions, some structure on the relation between the payoff matrices of the powerful and the powerless (Table 2) is therefore necessary.

We restrict worlds such that $\alpha > \delta > \beta, \gamma$. Under this restriction, the game played between the powerful has a particular contextual meaning. It is always good for the bystander that the powerful players choose the same action, where the outcome (X, X) is strictly better. In other words, the impact of powerful actions is a coordination problem for the bystander. In some sense then, we want moral players to achieve the outcome (X, X) and amoral players to achieve (Y, Y) , thus giving *group morality* its meaning endogenously. Second, in our world mis-coordinated actions are the worst for the bystander.

These restrictions however do not eliminate games with multiple pure strategy equilibria. Since our comparison is between fully informed players versus uninformed players, we take the stand that Harsanyi's purification is less compelling in a world of complete information than his (along with Reinhard Selten's) notion of risk dominance when it comes to equilibrium selection.⁴ Therefore, whenever there are multiple pure strategy equilibria, we use risk dominance (Harsanyi and Selten 1988) on π to select a pure strategy equilibrium on which we work. Clearly, therefore, our world excludes π -games like the Matching Pennies.

Fact 1 (see proof in Appendix) characterizes the strategic aspects of the world we study. There are a total of four possible scenarios wherein moral players arrive at the equilibrium profile (X, X) while amoral players arrive at (Y, Y) : (X, X) (or (Y, Y)) is the unique equilibrium in the game between moral (or amoral) players, or both (X, X) and (Y, Y) are equilibria with (X, X) (or (Y, Y)) being the risk dominant one for moral (or amoral) players. Fact 1 shows that for our specific choice of the order of bystander payoff parameters, only two of these scenarios are feasible.

Fact 1. *Under equilibrium selection via risk-dominance, if morals arrive at (X, X) and amorals arrive at (Y, Y) , and if $\alpha > \delta > \beta, \gamma$, then the payoff-matrix pair (A, B) has the following strategic properties:*

⁴This is because the purification-based refinement is after all a limit result when asymmetric information vanishes.

Table 3. Feasibility and Prisoner’s Dilemma

		Player 2	
		X	Y
Player 1	X	-1, -1	-3, 0
	Y	0, -3	-2, -2
		B_1	
		X	Y
	X	6	2
	Y	1	3
		B_2	
		X	Y
	X	3	1
	Y	2.5	2.75

- Feasibility:** (i) (Y, Y) is the unique equilibrium between amoral players. Both (X, X) and (Y, Y) are equilibria with two moral players, with (X, X) risk dominating (Y, Y); (ii) Both (X, X) and (Y, Y) are equilibria for games between two moral or two amoral players. Here (X, X) risk dominates (Y, Y) when agents are moral and (Y, Y) risk dominates (X, X) when agents are amoral.
- Infesibility:** (i) (X, X) is the unique equilibrium between moral players and (Y, Y) is the unique equilibrium between amoral players; (ii) (X, X) is the unique equilibrium between moral players whereas both (X, X) and (Y, Y) are equilibria with two amoral players, with (Y, Y) risk dominating (X, X).

Clearly, the restriction on matrix B gives us a universe of feasible games. This universe is unrelated to popular classifications such as the Prisoner’s Dilemma, the Battle of the Sexes, the Game of Chicken, pure coordination games or any other popular strategic situations. This implies that depending on the exact specifications of the pair of matrices (A, B) , some Prisoner’s Dilemmas or some Battle of the Sexes, etc. will fall in our feasible world while others will not. For example, consider Table 3 where the base game A is a Prisoner’s Dilemma with Y as the strictly dominant strategy and B_1 and B_2 are two bystander matrices both satisfying the condition $\alpha > \delta > \beta, \gamma$. However, using a linear m , we find that the pair (A, B_1) is feasible as in Fact 1 but the pair (A, B_2) is not.

4. Information and the Curse

Unlike in Basu’s example, when matched players in our world learn each other’s morality type before choosing their actions, the bystander’s payoff, when a moral and an amoral player are matched, is at least as much as when two amorals are matched. This follows, although tediously, from the fact that (X, X) and (Y, Y) are the only equilibria that can be realized for such a pair (see **A1** in Appendix). Hence, either the bystander gets a payoff of α or δ from the game. So, there can be no

possibility of a Samaritan's Curse. Proposition 1 shows that this holds as well for the bystander's aggregate/average welfare from the powerful population. The following Lemma is powerful and key to the proposition.

Lemma 1. *Consider a matching where one player is moral and the other player is amoral and types are common knowledge. Then, (X, X) and (Y, Y) are the only two equilibria irrespective of the position of these players.*

The proof of Lemma 1 is not straightforward and the details are provided in the Appendix. It shows that the construction of the universe of games that we focus on cannot yield mis-coordinated outcomes as equilibria irrespective of the types of the matched players. This simplifies the proof of Proposition 1.

Proposition 1. *Consider a world where powerful players learn each other's morals before they choose actions. Then for any warm-glow function m , the ex-ante expected utility of the bystander is increasing in μ , the proportion of moral players in the population.*

The proof of Proposition 1 is straightforward and we provide the steps directly here. Given Lemma 1, denote by T_Y the case that (Y, Y) is played in a match between a moral and an amoral player irrespective of which player is assigned the role of Player 1. Similarly, define T_X as the case that (X, X) is played in a match between a moral and an amoral player irrespective of which player is assigned the role of Player 1. Finally, let T_{XY} be the event where if a moral and an amoral player match then (Y, Y) is played in one assignment (irrespective of which player name is assigned to which player) and (X, X) is played in the other assignment. Then the expected payoff EU_B of the bystander before pairs are formed randomly is given by

$$EU_B = \begin{cases} \mu^2\alpha + 2\mu(1 - \mu)\delta + (1 - \mu)^2\delta, & \text{if } T_Y \\ \mu^2\alpha + \mu(1 - \mu)(\delta + \alpha) + (1 - \mu)^2\delta, & \text{if } T_{XY} \\ \mu^2\alpha + 2\mu(1 - \mu)\alpha + (1 - \mu)^2\delta, & \text{if } T_X \end{cases} \quad (2)$$

Note that the above cases enlist all possible scenarios wherein either (X, X) or (Y, Y) is realized when a moral and an amoral are chosen and are assigned the positions 1 and 2 or 2 and 1 respectively. For example, consider the first row of the expression for EU_B in (2). Here, we are in a strategic environment where whenever a pair is formed between a moral and an amoral player, each play Y irrespective of whether they are assigned the role of Player 1 or Player 2. Therefore, $2\mu(1 - \mu)$ proportion of matches are between a moral and an amoral player (of which half the cases are when the moral player takes the role of Player 1) but both players play Y (and the bystander earns δ), μ^2 proportion of matches are between two morals where each play X (and the bystander earns α), and $(1 - \mu)^2$ proportion of matches are between two amoral players where each play Y (and the bystander earns δ). The other two rows are obtained similarly.

To understand how EU_B changes as the proportion μ of morals rises in the population, we look at the first derivative of Equation (2), which gives the following:

$$\frac{\partial EU_B}{\partial \mu} = \begin{cases} 2\mu(\alpha - \delta), & \text{if } T_Y \\ \alpha - \delta, & \text{if } T_{XY} \\ 2(1 - \mu)(\alpha - \delta), & \text{if } T_X \end{cases} \quad (3)$$

It is straightforward to see that each of the above derivatives are greater than 0 for $\alpha > \delta$. Hence, we find that in a population where μ proportion are moral while the rest are amoral, the ex-ante expected utility of the bystander is strictly increasing in μ .

Remark: *It is important to note that the above result holds irrespective of whether the warm glow from moral consequences dampens or explodes with an increase in bystander welfare (that is irrespective of any further assumptions on the function $m(\cdot)$ beyond monotonicity). Such is the structure of the world that even if the marginal incentives between moral players and the bystander are not aligned (due to non-linearity), there is no perverse impact of morality when powerful players know each others' types. Hence, Fact 1 is sufficient for providing us with a universe of games (viz. the feasibility conditions therein) to answer the main question of this paper. It is important to note here that our analysis is towards a partial characterization of this question. There may exist other 2×2 games where immiserizing effects of morality on the welfare of the bystander are completely absent under complete information. One can attempt to write a general condition for this, but not without losing the structure we have in our world which helps us to obtain a clear theoretical comparison across the two information structures.*

We now shift our focus to the second informational structure wherein players' morality-types are not revealed after matching so that morality remains private information throughout the game.

The incomplete information game played between matched pairs of powerful players becomes a standard Bayesian game with private information. In such a game, a (Bayesian) strategy for an individual player is a complete plan of action for each type that the player can possibly be. This plan is a non-binding commitment device about what the player would choose before she knows her true type. The idea of a Bayes–Nash equilibrium that we employ in our analysis is that given other players are following their strategies, no player has a strict incentive to change their plan (read strategy) unilaterally. This property of 'equilibrium' is often called Incentive Compatibility, a term used below. These are defined (and constructed) formally in what follows.

Given the endogenous association of the outcome (X, X) with moral players, the following construction of a strategy-profile is without any loss of generality in the world of pure strategies:

Let $\sigma_i : \{\text{moral}, \text{amoral}\} \rightarrow \{X, Y\}$ be a Bayesian pure strategy, that maps Player i 's types to her actions, such that for each player $i \in \{1, 2\}$, $\sigma_i(\text{moral}) = X$ and $\sigma_i(\text{amoral}) = Y$. Hence, the plan is simply this: If the player is moral, s/he plays X while if s/he is amoral, s/he plays Y .

For this profile to be a Bayes–Nash equilibrium, the following are necessary and sufficient Incentive-Compatibility conditions. First consider Player 1 (the row player). In the event she is moral, and if she plays according to that plan, then, given Player 2 is playing according to the same plan, the payoff expected by Player 1 is $(a_1 + m(\alpha))\mu + (b_1 + m(\beta))(1 - \mu)$ since with probability μ Player 2 is moral and as she is playing according to the same plan, Player 1 expects the action X with probability μ (when Player 1 obtains $a_1 + m(\alpha)$) and Y with probability $1 - \mu$

(when Player 1 obtains $b_1 + m(\beta)$). If Player 1 contemplates to deviate, then the only deviation is to the action Y from which, for the same reason as above, she expects a payoff of $(c_1 + m(\gamma))\mu + (d_1 + m(\delta))(1 - \mu)$. Hence, Player 1 will not deviate unilaterally from her plan when she is moral if and only if

$$(a_1 + m(\alpha))\mu + (b_1 + m(\beta))(1 - \mu) \geq (c_1 + m(\gamma))\mu + (d_1 + m(\delta))(1 - \mu). \quad (4)$$

Using the same kind of reasoning, it then follows that Player 1 will not deviate unilaterally from her plan when she is amoral if and only if

$$c_1\mu + d_1(1 - \mu) \geq a_1\mu + b_1(1 - \mu) \quad (5)$$

Similarly, for player 2, they are

$$(a_2 + m(\alpha))\mu + (c_2 + m(\gamma))(1 - \mu) \geq (b_2 + m(\beta))\mu + (d_2 + m(\delta))(1 - \mu), \quad (6)$$

when Player 2 is moral and

$$b_2\mu + d_2(1 - \mu) \geq a_2\mu + c_2(1 - \mu), \quad (7)$$

when Player 2 is amoral. The strategy profile under study is a Bayes-Nash equilibrium if and only if the above four inequalities hold simultaneously, ensuring that no player in no role and in no type has an incentive to deviate unilaterally from the plan. In what follows, we will live in such an equilibrium, provided it exists, since existence is not guaranteed and depends on the parameters of the matrices (A, B) as well as the warm glow function $m(v(\cdot))$.

Whenever such an equilibrium where $\sigma = (\sigma_1, \sigma_2)$ as defined before exists, the expected payoff of the bystander is given by

$$EU_B(\sigma|\mu) = \mu^2\alpha + \mu(1 - \mu)\beta + (1 - \mu)\mu\gamma + (1 - \mu)^2\delta, \quad (8)$$

since with probability μ^2 both players are moral and we achieve (X, X) (where the bystander's payoff is α), with probability $(1 - \mu)^2$ both players are amoral and we achieve (Y, Y) (where the bystander's payoff is δ), with probability $\mu(1 - \mu)$ only Player 1 is moral and so we achieve (X, Y) (where the bystander's payoff is β), and finally with probability $(1 - \mu)\mu$ only Player 1 is amoral and so we achieve (Y, X) (where the bystander's payoff is γ).

We are interested in scenarios wherein even though the proportion of moral agents μ increases, the expected payoff of the bystander decreases. Hence our focus must be on the conditions enlisted in Lemma 2.⁵

⁵*Games with strictly dominant strategies for all players:* For any pair of payoff matrices (A, B) , if both morals and amorals have their respective and distinct (strictly) dominant strategies (note: such a game is not feasible given Fact 1), information about the opponent's type does not affect the outcome of the game. In any such game, a rise in μ increases the bystander's payoff. To see this, notice that if Player i is moral, we need $a_i + m(\alpha) > b_i + m(\beta)$, $c_i + m(\gamma)$ and at the same time, $a_i < b_i$, c_i . Hence, $\alpha > \beta$, γ . Similarly, we get the relation $\beta, \gamma > \delta$. For these conditions, equations (9) and (10) will fail to hold for any $\mu \in (0, 1)$.

Table 4. Samaritan’s Curse under incomplete information

		Player 2		
		X	Y	
Player 1	X	70, 34	8, 5	
	Y	19, -1	81, 40	
		Bystander		
		X	Y	
		X	9	2
		Y	1	6

Lemma 2. $\partial EU_B / \partial \mu < 0$ if and only if

$$\mu < \frac{\delta - \frac{\beta + \gamma}{2}}{(\alpha + \delta) - (\beta + \gamma)} \text{ when } \alpha + \delta > \beta + \gamma \tag{9}$$

$$\text{and } \mu > \frac{\frac{\beta + \gamma}{2} - \delta}{(\beta + \gamma) - (\alpha + \delta)} \text{ when } \alpha + \delta < \beta + \gamma \tag{10}$$

The challenging question is therefore: are there robust specifications of the parameter space such that Lemma 2 and the four Incentive Compatibility conditions hold simultaneously. Proposition 2 shows that there exist conditions on the game parameters and the warm-glow function m such that the answer is yes.

Before moving to the proposition, we illustrate this further through an example. Consider the game matrix A and bystander’s payoff matrix B in Table 4. Let $m(\cdot)$ be the convex function $m(v) = v^3$ for $v > 0$.⁶ Then, σ is a Bayes Nash equilibrium for the above game for any population where the proportion of morals $\mu \in (0.265, 0.585)$. However, expected payoff of the bystander is decreasing in μ for $\mu < 0.375$. Hence, for $\mu \in (0.265, 0.375)$, the expected payoff of the bystander is decreasing in the proportion of moral agents in the population.

Proposition 2 is a powerful result within the world of strategic environments described by Fact 1 since it shows that for *any* environment in that world, the Samaritan’s curse is obtainable solely due to asymmetric information, if and only if the warm glow function is sufficiently convex (viz. euphoric).

⁶In light of Fact 1, note that for any morality profile of a given match, both (X, X) and (Y, Y) are equilibrium points but (X, X) risk dominates if and only if both players are moral, and otherwise the risk dominant equilibrium is (Y, Y) . For intellectual curiosity, it is true that in this example using even a linear m , the bystander’s payoff is maximum when both players are amoral if we move to mixed strategies where actions lose the endogenous meaning of morality, discussed in the paragraphs before Fact 1. Note that, this conundrum of mixed strategies is absent in Basu’s example (viz. Table 2) since equilibria are unique and in pure strategies.

Proposition 2. *Suppose morality types are private information. Then there exists a non-empty interval of μ denoted by $(\underline{\mu}, \bar{\mu}) \subset [0, 1]$ such that (i) $\bar{\mu}$ depends only on the parameters of the matrices (A, B) while $\underline{\mu}$ depends both on the parameters of the matrices (A, B) and the nature of the warm-glow function $m(v(\cdot))$ and (ii) for any $\mu \in (\underline{\mu}, \bar{\mu})$, σ is a Bayes Nash equilibrium with $\partial EU_{B|\sigma} / \partial \mu < 0$ if and only if $m(\cdot)$ exhibits sufficient euphoria.*

In the Appendix, we report the exact values of the two bounds for μ as stated in Proposition 2 (see (78)). They are

$$\underline{\mu} = \max \left\{ \frac{d_1 + m(\delta) - b_1 - m(\beta)}{a_1 + m(\alpha) - b_1 - m(\beta) - c_1 - m(\gamma) + d_1 + m(\delta)}, \frac{d_2 + m(\delta) - c_2 - m(\gamma)}{a_2 + m(\alpha) - b_2 - m(\beta) - c_2 - m(\gamma) + d_2 + m(\delta)} \right\},$$

and

$$\bar{\mu} = \min \left\{ \frac{d_1 - b_1}{a_1 - b_1 - c_1 + d_1}, \frac{d_2 - c_2}{a_2 - b_2 - c_2 + d_2}, \frac{\delta - \frac{\beta + \gamma}{2}}{(\alpha + \delta) - (\beta + \gamma)} \right\}$$

As seen in the proof, given a sufficiently convex warm-glow function m , at any μ_0 in the interval $(\underline{\mu}, \bar{\mu})$ which is non-empty, σ is a Bayes Nash equilibrium and the bystander’s utility is decreasing as the proportion of morality increases from μ . This interval becomes larger as the convexity of the warm-glow function increases. Moreover, reducing convexity of m destroys this result; for example, when m is linear, the result is not true for sure, and therefore not true for concave warm glow functions either.

We illustrate the findings in Proposition 2 by looking at an example. Consider the game given in Table 4 and consider the warm-glow function $m(x) = x^t, x \in [0, \infty)$, for $t > 1$, the degree of euphoria. Clearly, the convexity of m increases in t . Figure 1 shows how the upper bound and lower bound of the aforementioned interval in (78) are affected by the convexity of m . As t increases, the upper bound of μ_0 remains constant while the lower bound is decreasing in t in the range $(1, \infty)$. As can be clearly seen from the plot, this feasible interval (viz. the shaded region) for the Samaritan’s Curse is empty for $t \leq 2.35$ approximately. Hence, for this example, m needs to be at least as convex as $x^{2.4}$ to obtain the Samaritan’s curse.

Underlying intuition: The intuition behind the contrast between propositions 1 and 2 is rooted in a number of aspects of the environment studied. The first is our fundamental premise, absent in Basu’s example, that observable action profiles have their endogenous contextual meanings when it comes to group morality. In particular, observing the outcome (X, X) will imply that at least one player is moral and observing (Y, Y) will imply that at least one player is amoral. In our world specified by Fact 1 and as shown in the proof of proposition 1, when types are common knowledge, outcomes like (X, Y) and (Y, X) are never obtained in any equilibrium irrespective of the morality profile of the matched pairs or the exact nature of the warm glow function.

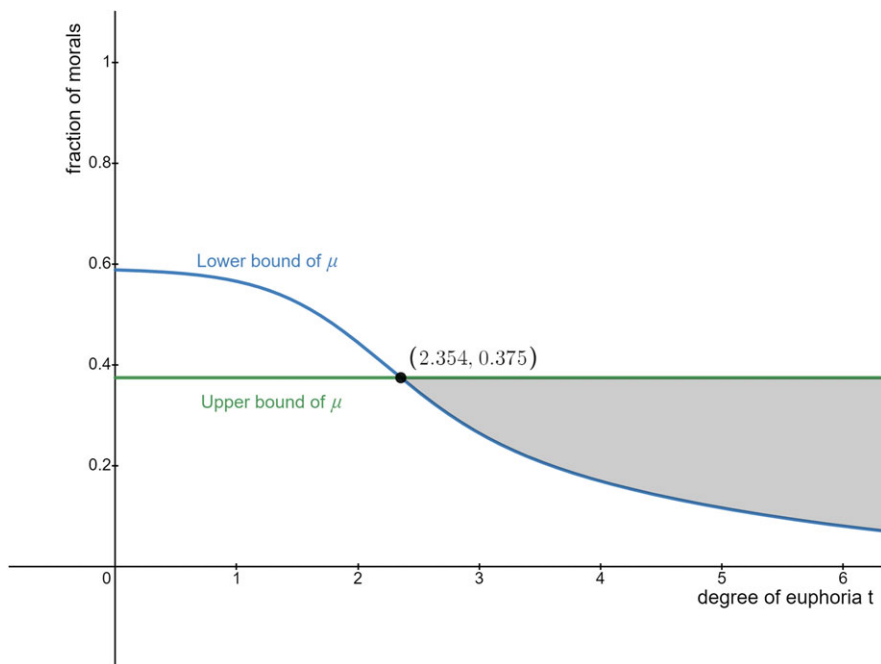


Figure 1. Upper bound vs Lower bound plot of μ_0 .

On the other hand, in the same world, incomplete information strengthens the contextual meaning of the action X naturally whereby in equilibrium, X is played if and only if the player is moral while Y is played if and only if the player is amoral. This yields positive probabilities on the outcomes (X, Y) and (Y, X) when the matched players are morally different. We then show that despite the enlargement of the outcome space that allows for the worst bystander outcomes (*viz.* β, γ), sufficient euphoria (sufficiently convex m) is necessary to obtain a Samaritan's curse-like phenomenon. This is because it is only when m is sufficiently convex that moral players, facing uncertainty about the morality of their partners are willing to take the risk of playing X , the moral action, knowing fully well that if their partner was amoral the bystander is worst off. Since the bystander's payoff from matrix B is B itself, the bystander himself does not want this risk to be taken. This mismatch in the intensity of incentives between the powerful morals and the bystander himself implies that for certain values of μ , a rise in μ distorts the bystander's interests, thereby generating the Samaritan's curse. Of course, when μ is larger than the upper bound (see Figure 1 for an example) there is hardly any risk while when μ is lower than the lower bound, there is hardly anyone to take this risk.

4.1. Culture and Social Connectivity

One way of interpreting the social context of the two main propositions is to consider and contrast societies (or organizations in general) that are on one hand more or less concerned about the powerless (captured by μ), while on the other,

they are culturally different so that in one, individuals are more connected and know each other, while in the other the individuals are less connected. Cultural differences can play a major role in aggregate activity in organizations and countries as demonstrated in the pioneering works of Trompenaars (1993) and Hofstede (2001). Cultural distinction becomes relevant in the discourse of the present paper because it is more likely that in culturally more individualistic societies where communication is scarce, individual moral codes are not publicly observed. Thus Proposition 1 is a better representation of connected societies while Proposition 2 captures this essence of a highly disconnected one.

From this point of view, we see that in a universe of non-cooperative strategic environments (characterized by Fact 1), a rise in the number of moral individuals in a less connected society unambiguously enhances the society's act-morality from a consequentialist position. This cannot be said universally when we consider highly disconnected societies, particularly those where the proportion of moral individuals is neither too large nor too small. If one must take public campaigns on caring for the powerless literally, it also means that such campaigns are always beneficial for connected societies while one needs to be careful in those which are largely disconnected.

5. Conclusion

This paper provides a partial characterization of strategic environments where the surprising phenomenon called the Samaritan's Curse introduced by Basu (2022) is driven solely due to lack of common knowledge about morality amongst powerful players. We introduce the notion of *euphoria* and highlight the dependence of the curse on the level of euphoria moral players exhibit. Sufficiently euphoric moral players become so zealous about bystander payoff that they choose a bystander-benevolent action even if the associative risk of mis-coordination is so great that the bystander himself isn't willing to take that risk. This leads to the Samaritan's curse when information about morality is withheld from powerful players. Further, we also find that there exist environments outside our world wherein such phenomena are not possible irrespective of the information these players have. The findings of this paper are very important for policy design – for example, public campaigns on caring for the powerless would benefit culturally close-knit communities while having the opposite effect in communities that are rather disconnected. Moreover, the results obtained here can be applied in spheres of public shaming of immoral actions or when powerful players are also spiteful with respect to the welfare of the powerless.

The environment introduced by Basu (2022) has similarities with Dictator Games but it is not necessarily zero-sum and there are multiple dictators such that the final 'offer' they, as a group, make to the powerless bystanders are outcomes of strategic games played between these dictators with conflicting interests. Such a world is vast and complex, and several important avenues of future research can stem from here. Studying this carefully is essential for our understanding of problems relating to environment, climate and global poverty.

Acknowledgements. We thank Kaushik Basu for introducing us to the fascinating world of aggregation of morality through various insightful discussions. We also thank Debabrata Datta, Saptarshi P. Ghosh,

Rajnish Kumar, K. S. Mallikarjuna Rao, Diksha Singh and Grazyna Wiejak-Roy for various comments and suggestions. As part of a larger project on group morality, this paper has also benefited from presentations at University of Warsaw, IIT Mumbai, IIFT Kolkata, Krea University and Queen's University Belfast.

References

- Basu K.** 2010. The moral basis of prosperity and oppression: altruism, other-regarding behaviour and identity. *Economics & Philosophy* **26**, 189–216.
- Basu K.** 2022. The samaritan's curse: moral individuals and immoral groups. *Economics & Philosophy* **38**, 132–151.
- Camerer C.F. and R.H. Thaler** 1995. Anomalies: ultimatums, dictators and manners. *Journal of Economic Perspectives* **9**, 209–219.
- Cooper D.J. and J.H. Kagel** 2016. A failure to communicate: an experimental investigation of the effects of advice on strategic play. *European Economic Review* **82**, 24–45.
- Engel C.** 2011. Dictator games: a meta study. *Experimental Economics* **14**, 583–610.
- Fehr E. and S. Gächter** 2002. Altruistic punishment in humans. *Nature* **415**(6868), 137–140.
- Fehr E. and K. M. Schmidt** 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* **114**, 817–868.
- Fehr E. and K. M. Schmidt** 2006. The economics of fairness, reciprocity and altruism – experimental evidence and new theories. In *Foundations. Volume 1, Handbook of the Economics of Giving, Altruism and Reciprocity*, ed. S.-C. Kolm and J. Mercier Ythier, 615–691. New York: Elsevier.
- Harsanyi J.C. and R. Selten** 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.
- Henrich J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis and R. McElreath.** 2001. In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review* **91**, 73–78.
- Hofstede G.** 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. 2nd edn. Thousand Oaks: Sage.
- Hume D.** 1740. *A Treatise of Human Nature*.
- Kahneman D., J. L. Knetsch and R. Thaler.** 1986. Fairness as a constraint on profit seeking: entitlements in the market. *American Economic Review* **76**(4), 728–741.
- Kant I.** 1764. *Inquiry concerning the Distinctness of the Principles of Natural Theology and Morality*.
- Schwartz S.** 1996. Value priorities and behavior: applying a theory of integrated value systems. In *The Psychology of Values: The Ontario Symposium*, Vol. 8, ed. C. Seligman, J.M. Olson and M.P. Zanna, 1–24. New York: Lawrence Erlbaum Associates.
- Trompenaars F.** 1993. *Riding the Waves of Culture: Understanding Cultural Diversity in Global Business*. London: McGraw-Hill.

Appendix

Proof of Fact 1

Proof. Consider 2(i) of Fact 1 where both moral agents and amoral agents have unique equilibria in their respective games. Then, the following conditions must hold for morals and amorals respectively:

$$a_1 + m(\alpha) > c_1 + m(\gamma) \text{ and } a_2 + m(\alpha) > b_2 + m(\beta) \quad (11)$$

$$d_1 + m(\delta) < b_1 + m(\beta) \text{ or } d_2 + m(\delta) < c_2 + m(\gamma) \quad (12)$$

$$d_1 > b_1 \text{ and } d_2 > c_2 \quad (13)$$

$$a_1 < c_1 \text{ or } a_2 < b_2 \quad (14)$$

Without loss of generality, let $d_1 + m(\delta) < b_1 + m(\beta)$ from Equation (12). But $d_1 > b_1$ and according to our assumption $\delta > \beta$. This leads to a contradiction. Hence, this scenario is not possible.

Next for 2(ii), we have the following conditions:

$$a_1 + m(\alpha) > c_1 + m(\gamma) \text{ and } a_2 + m(\alpha) > b_2 + m(\beta) \tag{15}$$

$$d_1 + m(\delta) < b_1 + m(\beta) \text{ or } d_2 + m(\delta) < c_2 + m(\gamma) \tag{16}$$

$$d_1 > b_1 \text{ and } d_2 > c_2 \tag{17}$$

$$a_1 > c_1 \text{ and } a_2 > b_2 \tag{18}$$

$$(b_1 - d_1)(c_2 - d_2) > (c_1 - a_1)(b_2 - a_2) \tag{19}$$

Similar to scenario (a), Equations (16) and (17) together form a contradiction. Hence, we see that for $\delta > \beta, \gamma$ if (Y, Y) is an equilibrium between amorals, (X, X) cannot be the unique equilibrium in the game between two morals. \square

A1

1. Consider the first scenario wherein (Y, Y) is the unique equilibrium between two amoral agents. Further, (X, X) is the realized equilibrium for the game between moral agents (i.e. (X, X) risk-dominates (Y, Y)). The following conditions must then hold.

$$a_1 + m(\alpha) > c_1 + m(\gamma) \text{ and } a_2 + m(\alpha) > b_2 + m(\beta) \tag{20}$$

$$d_1 + m(\delta) > b_1 + m(\beta) \text{ and } d_2 + m(\delta) > c_2 + m(\gamma) \tag{21}$$

$$(c_1 + m(\gamma) - a_1 - m(\alpha))(b_2 + m(\beta) - a_2 - m(\alpha)) > (b_1 + m(\beta) - d_1 - m(\delta))(c_2 + m(\gamma) - d_2 - m(\delta)) \tag{22}$$

$$d_1 > b_1 \text{ and } d_2 > c_2 \tag{23}$$

$$a_1 < c_1 \text{ or } a_2 < b_2 \tag{24}$$

We further look at equilibria in games between a moral and an amoral player when the above conditions 33 through 37 hold. If the moral player is assigned to be Player 1, then we can say the following about the ensuing equilibria:

(a) If $b_2 > a_2$, (Y, Y) is the unique equilibrium.

(b) If $b_2 < a_2$, both (X, X) and (Y, Y) are possible equilibria, with (Y, Y) being realized if

$$(c_1 + m(\gamma) - a_1 - m(\alpha))(b_2 - a_2) < (b_1 + m(\beta) - d_1 - m(\delta))(c_2 - d_2) \tag{25}$$

that is, (Y, Y) risk dominates (X, X) .

If the above conditions do not hold, (X, X) is realized.

However if the moral player is assigned to be Player 2, the conditions for which (Y, Y) is realized change, as given below:

(a) If $c_1 > a_1$, (Y, Y) is the unique equilibrium.

(b) If $c_1 < a_1$, both (X, X) and (Y, Y) are possible equilibria, with (Y, Y) being realized if

$$(c_1 - a_1)(b_2 + m(\beta) - a_2 - m(\alpha)) < (b_1 - d_1)(c_2 + m(\gamma) - d_2 - m(\delta)) \quad (26)$$

that is, (Y, Y) risk dominates (X, X) .

Similar to before, if the above conditions do not hold, (X, X) is realized.

2. Next, consider the scenario when both (X, X) and (Y, Y) are equilibria when morals ((X, X) risk dominates) or amorals ((Y, Y) risk dominates) meet. In this scenario, the following conditions hold,

$$a_1 + m(\alpha) > c_1 + m(\gamma) \text{ and } a_2 + m(\alpha) > b_2 + m(\beta) \quad (27)$$

$$d_1 + m(\delta) > b_1 + m(\beta) \text{ and } d_2 + m(\delta) > c_2 + m(\gamma) \quad (28)$$

$$(c_1 + m(\gamma) - a_1 - m(\alpha))(b_2 + m(\beta) - a_2 - m(\alpha)) > (b_1 + m(\beta) - d_1 - m(\delta))(c_2 + m(\gamma) - d_2 - m(\delta)) \quad (29)$$

$$d_1 > b_1 \text{ and } d_2 > c_2 \quad (30)$$

$$a_1 > c_1 \text{ and } a_2 > b_2 \quad (31)$$

$$(b_1 - d_1)(c_2 - d_2) > (c_1 - a_1)(b_2 - a_2) \quad (32)$$

Here, both (X, X) and (Y, Y) are equilibria of the game between a moral and an amoral player, with the risk-dominant equilibrium being realized.

Proof of Lemma 1

Proof. Consider the first feasible scenario wherein (Y, Y) is the unique equilibrium between two amoral agents. Further, (X, X) is the realized equilibrium for the game between moral agents (i.e., (X, X) risk-dominates (Y, Y)). The following conditions must then hold.

$$a_1 + m(\alpha) > c_1 + m(\gamma) \text{ and } a_2 + m(\alpha) > b_2 + m(\beta) \quad (33)$$

$$d_1 + m(\delta) > b_1 + m(\beta) \text{ and } d_2 + m(\delta) > c_2 + m(\gamma) \quad (34)$$

$$(c_1 + m(\gamma) - a_1 - m(\alpha))(b_2 + m(\beta) - a_2 - m(\alpha)) > (b_1 + m(\beta) - d_1 - m(\delta))(c_2 + m(\gamma) - d_2 - m(\delta)) \quad (35)$$

$$d_1 > b_1 \text{ and } d_2 > c_2 \quad (36)$$

$$a_1 < c_1 \text{ or } a_2 < b_2 \quad (37)$$

We now look at equilibria in games between a moral and an amoral player when the above conditions 33 through 37 hold. If the moral player is assigned to be Player 1, then we can say the following about the ensuing equilibria:

1. If $b_2 > a_2$, (Y, Y) is the unique equilibrium.
2. If $b_2 < a_2$, both (X, X) and (Y, Y) are possible equilibria, with (Y, Y) being realized if

$$(c_1 + m(\gamma) - a_1 - m(\alpha))(b_2 - a_2) < (b_1 + m(\beta) - d_1 - m(\delta))(c_2 - d_2) \quad (38)$$

that is, (Y, Y) risk dominates (X, X) .

If the above conditions do not hold, (X, X) is realized.

However if the moral player is assigned to be Player 2, the conditions for which (Y, Y) is realized change, as given below:

1. If $c_1 > a_1$, (Y, Y) is the unique equilibrium.
2. If $c_1 < a_1$, both (X, X) and (Y, Y) are possible equilibria, with (Y, Y) being realized if

$$(c_1 - a_1)(b_2 + m(\beta) - a_2 - m(\alpha)) < (b_1 - d_1)(c_2 + m(\gamma) - d_2 - m(\delta)) \quad (39)$$

that is, (Y, Y) risk dominates (X, X) .

Similar to before, if the above conditions do not hold, (X, X) is realized.

Next, consider the scenario when both (X, X) and (Y, Y) are equilibria when morals ((X, X) risk dominates) or amorals ((Y, Y) risk dominates) meet. In this scenario, the following conditions hold,

$$a_1 + m(\alpha) > c_1 + m(\gamma) \text{ and } a_2 + m(\alpha) > b_2 + m(\beta) \quad (40)$$

$$d_1 + m(\delta) > b_1 + m(\beta) \text{ and } d_2 + m(\delta) > c_2 + m(\gamma) \quad (41)$$

$$(c_1 + m(\gamma) - a_1 - m(\alpha))(b_2 + m(\beta) - a_2 - m(\alpha)) > (b_1 + m(\beta) - d_1 - m(\delta))(c_2 + m(\gamma) - d_2 - m(\delta)) \quad (42)$$

$$d_1 > b_1 \text{ and } d_2 > c_2 \quad (43)$$

$$a_1 > c_1 \text{ and } a_2 > b_2 \quad (44)$$

$$(b_1 - d_1)(c_2 - d_2) > (c_1 - a_1)(b_2 - a_2) \quad (45)$$

Here, both (X, X) and (Y, Y) are equilibria of the game between a moral and an amoral player, with the risk-dominant equilibrium being realized. □

Proof of Proposition 2

Proof. The proof consists of listing conditions which need to be satisfied for a μ to exist such that σ is a Bayes Nash equilibrium and expected utility of the bystander is decreasing in μ . We show that for any linear warm glow function, these conditions cannot simultaneously hold and hence, the expected utility of the bystander is always increasing in μ . We then proceed to show that when m is sufficiently convex, there exists an interval such that any μ in this interval satisfies all these conditions, leading to the result.

We first consider part 1.(ii) of Fact 1. Let Equations (40) to (45) hold. Moreover, let Equations (4) to (7) hold for σ to be a Bayes Nash equilibrium. Given the aforementioned conditions on our base game and bystander payoff matrix

parameters, Equations (4) to (7) boil down to the following bounds on μ , in that order.

$$\mu > \frac{d_1 + m(\delta) - b_1 - m(\beta)}{a_1 + m(\alpha) - b_1 - m(\beta) - c_1 - m(\gamma) + d_1 + m(\delta)} \tag{46}$$

$$\mu < \frac{d_1 - b_1}{a_1 - b_1 - c_1 + d_1} \tag{47}$$

$$\mu > \frac{d_2 + m(\delta) - c_2 - m(\gamma)}{a_2 + m(\alpha) - b_2 - m(\beta) - c_2 - m(\gamma) + d_2 + m(\delta)} \tag{48}$$

$$\mu < \frac{d_2 - c_2}{a_2 - b_2 - c_2 + d_2} \tag{49}$$

These conditions give feasibility conditions on μ for σ to be a Bayes Nash Equilibrium. Moreover, each of these bounds lie in $(0, 1)$. Further, from Equation (9), we get an upper bound of μ below which the Expected Utility of the bystander decreases as μ increases. We intend to show that a μ and an interval I around it exist such that as μ increases in I , the expected utility of the bystander decreases. Using the aforementioned equations, we find conditions which are necessary and sufficient for the existence of such a μ , in this framework.

$$\frac{d_1 + m(\delta) - b_1 - m(\beta)}{a_1 + m(\alpha) - b_1 - m(\beta) - c_1 - m(\gamma) + d_1 + m(\delta)} < \frac{\delta - \frac{\beta + \gamma}{2}}{(\alpha + \delta) - (\beta + \gamma)} \tag{50}$$

$$\frac{d_2 + m(\delta) - c_2 - m(\gamma)}{a_2 + m(\alpha) - b_2 - m(\beta) - c_2 - m(\gamma) + d_2 + m(\delta)} < \frac{\delta - \frac{\beta + \gamma}{2}}{(\alpha + \delta) - (\beta + \gamma)} \tag{51}$$

$$\frac{d_1 + m(\delta) - b_1 - m(\beta)}{a_1 + m(\alpha) - b_1 - m(\beta) - c_1 - m(\gamma) + d_1 + m(\delta)} < \frac{d_1 - b_1}{a_1 - b_1 - c_1 + d_1} \tag{52}$$

$$\frac{d_2 + m(\delta) - c_2 - m(\gamma)}{a_2 + m(\alpha) - b_2 - m(\beta) - c_2 - m(\gamma) + d_2 + m(\delta)} < \frac{d_2 - c_2}{a_2 - b_2 - c_2 + d_2} \tag{53}$$

Equations (50) and (51) ensure that there exists μ for which expected utility of the bystander is decreasing. Equations (52) and (53) ensure the non-emptiness of the feasible region of μ for which σ is a Bayes Nash equilibrium. Consider $m(\cdot)$ to be a linear function, i.e. $m(x) = sx + l$. As m is monotone and increasing, s has to be positive. We show that the above four conditions cannot simultaneously hold for a linear warm-glow function m .

As all numerators and denominators involved are positive values, Equation (50) can also be written in the following form.

$$(d_1 - b_1)(\alpha + \delta - \beta - \gamma) + (m(\delta) - m(\beta))(\alpha + \delta - \beta - \gamma) <$$

$$\left(\delta - \frac{\beta + \gamma}{2}\right)(a_1 - b_1 - c_1 + d_1) + \left(\delta - \frac{\beta + \gamma}{2}\right)(m(\alpha) - m(\beta) - m(\gamma) + m(\delta)) \tag{54}$$

For $k_1 = (d_1 - b_1)(\alpha + \delta - \beta - \gamma)$ and $k_2 = \left(\delta - \frac{\beta + \gamma}{2}\right)(a_1 - b_1 - c_1 + d_1)$, we have the following form of the above equation,

$$\left(\delta - \frac{\beta + \gamma}{2}\right)(m(\alpha) - m(\beta) - m(\gamma) + m(\delta)) - (m(\delta) - m(\beta))(\alpha + \delta - \beta - \gamma) > k_1 - k_2 \tag{55}$$

Similarly, Equation (51) can be written as,

$$\left(\delta - \frac{\beta + \gamma}{2}\right)(m(\alpha) - m(\beta) - m(\gamma) + m(\delta)) - (m(\delta) - m(\gamma))(\alpha + \delta - \beta - \gamma) > k_3 - k_4 \tag{56}$$

where $k_3 = (d_2 - c_2)(\alpha + \delta - \beta - \gamma)$ and $k_4 = \left(\delta - \frac{\beta + \gamma}{2}\right)(a_2 - b_2 - c_2 + d_2)$. Some simple algebraic calculations reduce the RHS of the above two equations to the following form:

$$k_1 - k_2 = \left(\alpha - \frac{\beta + \gamma}{2}\right)(d_1 - b_1) - \left(\delta - \frac{\beta + \gamma}{2}\right)(a_1 - c_1) \tag{57}$$

$$k_3 - k_4 = \left(\alpha - \frac{\beta + \gamma}{2}\right)(d_2 - c_2) - \left(\delta - \frac{\beta + \gamma}{2}\right)(a_2 - b_2) \tag{58}$$

Next, consider Equation (52) and (53). Expanding and reducing these equations yield the following:

$$(a_1 - c_1)(m(\delta) - m(\beta)) < (d_1 - b_1)(m(\alpha) - m(\gamma)) \tag{59}$$

$$(a_2 - b_2)(m(\delta) - m(\gamma)) < (d_2 - c_2)(m(\alpha) - m(\beta)) \tag{60}$$

Substituting $m(x) = sx + l$ in the above equations, we get

$$(a_1 - c_1)(\delta - \beta) < (d_1 - b_1)(\alpha - \gamma) \tag{61}$$

$$(a_2 - b_2)(\delta - \gamma) < (d_2 - c_2)(\alpha - \beta) \tag{62}$$

Equation (61) can be written as

$$\begin{aligned} (d_1 - b_1)\left(\alpha - \frac{\gamma + \beta}{2} - \frac{\gamma - \beta}{2}\right) &> (a_1 - c_1)\left(\delta - \frac{\beta + \gamma}{2} - \frac{\beta - \gamma}{2}\right) \Leftrightarrow \\ \left(\alpha - \frac{\beta + \gamma}{2}\right)(d_1 - b_1) - \left(\delta - \frac{\beta + \gamma}{2}\right)(a_1 - c_1) &> (d_1 - b_1 + a_1 - c_1)\left(\frac{\gamma - \beta}{2}\right) \end{aligned} \tag{63}$$

Similarly, Equation (62) can be written as

$$\left(\alpha - \frac{\beta + \gamma}{2}\right)(d_2 - c_2) - \left(\delta - \frac{\beta + \gamma}{2}\right)(a_2 - b_2) > (d_2 - b_2 + a_2 - c_2)\left(\frac{\beta - \gamma}{2}\right) \tag{64}$$

Without loss of generality, let $\beta \geq \gamma$. Then the LHS of Equation (64) is positive. This implies that $k_3 - k_4 > 0$ from Equation (58). Now, consider the LHS of Equation (56) and substitute $m(\cdot)$ as before. We have,

$$\begin{aligned} & \left(\delta - \frac{\beta + \gamma}{2}\right)(m(\alpha) - m(\beta) - m(\gamma) + m(\delta)) - (m(\delta) - m(\gamma))(\alpha + \delta - \beta - \gamma) \\ &= \left(\delta - \frac{\beta + \gamma}{2}\right)s(\alpha - \beta - \gamma + \delta) - s(\delta - \gamma)(\alpha + \delta - \beta - \gamma) \\ &= s\left(\frac{\gamma - \beta}{2}\right)(\alpha - \beta - \gamma + \delta) \leq 0 \end{aligned} \tag{65}$$

From the above we observe that the LHS of Equation (56) is non-positive while the RHS is positive. This leads to a contradiction to Equation (56). Hence, a linear warm-glow function m cannot lead to the Samaritan’s Curse.

Next, consider part 1.(i) of Fact 1. We show that if m is linear, Samaritan’s Curse cannot happen in this scenario as well. This scenario boils down to three sub-scenarios: (i) $a_1 < c_1$ and $a_2 > b_2$, (ii) $a_1 > c_1$ and $a_2 < b_2$ and (iii) $a_1 < c_1$ and $a_2 < b_2$. Consider the first case where $a_1 < c_1$ and $a_2 > b_2$. Here, again two sub-cases arise: (a) $a_1 + d_1 > b_1 + c_1$ and (b) $a_1 + d_1 < b_1 + c_1$. If $a_1 + d_1 > b_1 + c_1$, Equations (46) to (49) remain the same. Hence, Equations (50) to (53) need to hold for Samaritan’s Curse to take place. The rest of the proof follows by similar arguments as that in part 1.(ii) above. If however $a_1 + d_1 < b_1 + c_1$, Equations (46), (48) and (49) remain the same, while Equation (47) becomes:

$$\mu > \frac{b_1 - d_1}{b_1 + c_1 - a_1 - d_1} \tag{66}$$

However as $d_1 > b_1$, $\frac{b_1 - d_1}{b_1 + c_1 - a_1 - d_1} < 0$ and with $\mu \in (0, 1)$, the above condition is trivially satisfied. Hence, for Samaritan’s Curse to happen, only Equations (50), (51) and (53) need to simultaneously hold. Equation (50) can be written as,

$$\frac{m(\alpha) - m(\gamma) + a_1 - c_1}{m(\delta) - m(\beta) + d_1 - b_1} > \frac{\alpha - \frac{\beta + \gamma}{2}}{\delta - \frac{\beta + \gamma}{2}} \tag{67}$$

For a linear m , this equation becomes,

$$\frac{s(\alpha - \gamma) + a_1 - c_1}{s(\delta - \beta) + d_1 - b_1} > \frac{\alpha - \frac{\beta + \gamma}{2}}{\delta - \frac{\beta + \gamma}{2}} \tag{68}$$

As $a_1 - c_1 < 0$,

$$\frac{\alpha - \gamma}{\delta - \beta} = \frac{s(\alpha - \gamma)}{s(\delta - \beta)} > \frac{s(\alpha - \gamma) + a_1 - c_1}{s(\delta - \beta) + d_1 - b_1} > \frac{\alpha - \frac{\beta + \gamma}{2}}{\delta - \frac{\beta + \gamma}{2}} \tag{69}$$

Similarly, for the second player, we get

$$\frac{s(\alpha - \beta) + a_2 - b_2}{s(\delta - \gamma) + d_2 - c_2} > \frac{\alpha - \frac{\beta + \gamma}{2}}{\delta - \frac{\beta + \gamma}{2}} \tag{70}$$

Next, using the same linear form of m in Equation (53), we get

$$\frac{\alpha - \beta}{\delta - \gamma} > \frac{a_2 - b_2}{d_2 - c_2} \tag{71}$$

Therefore,

$$\frac{\alpha - \beta}{\delta - \gamma} = \frac{(s + 1)(\alpha - \beta)}{(s + 1)(\delta - \gamma)} > \frac{s(\alpha - \beta) + a_2 - b_2}{s(\delta - \gamma) + d_2 - c_2} > \frac{\alpha - \frac{\beta + \gamma}{2}}{\delta - \frac{\beta + \gamma}{2}} \tag{72}$$

Equations (69) and (72) together form a contradiction as $\frac{\alpha - \beta}{\delta - \gamma}$ and $\frac{\alpha - \gamma}{\delta - \beta}$ cannot both be greater than $\frac{\alpha - \frac{\beta + \gamma}{2}}{\delta - \frac{\beta + \gamma}{2}}$. Hence, the Samaritan’s Curse is not possible for this sub-case either when m is linear. The proofs of Samaritan’s Curse not being possible when m is linear under sub-scenarios (ii) and (iii) follow similar arguments.

Next, consider m to be a convex function. As before, we begin by considering Scenario 2. As already seen before, Equation (54) reduces to the following form:

$$\begin{aligned} & \left(\delta - \frac{\beta + \gamma}{2}\right)(m(\alpha) - m(\gamma)) - \left(\alpha - \frac{\beta + \gamma}{2}\right)(m(\delta) - m(\beta)) > \\ & \left(\alpha - \frac{\beta + \gamma}{2}\right)(d_1 - b_1) - \left(\delta - \frac{\beta + \gamma}{2}\right)(a_1 - c_1) \\ \Leftrightarrow & \frac{m(\alpha) - m(\gamma) + a_1 - c_1}{m(\delta) - m(\beta) + d_1 - b_1} > \frac{\alpha - \frac{\beta + \gamma}{2}}{\delta - \frac{\beta + \gamma}{2}} \end{aligned} \tag{73}$$

Similarly, for the second player, Equation (51) becomes,

$$\frac{m(\alpha) - m(\beta) + a_2 - b_2}{m(\delta) - m(\gamma) + d_2 - c_2} > \frac{\alpha - \frac{\beta + \gamma}{2}}{\delta - \frac{\beta + \gamma}{2}} \tag{74}$$

Next, consider Equations (52) and (53). Expanding and reducing these equations yield the following:

$$\frac{m(\alpha) - m(\gamma)}{m(\delta) - m(\beta)} > \frac{a_1 - c_1}{d_1 - b_1} \tag{75}$$

$$\frac{m(\alpha) - m(\beta)}{m(\delta) - m(\gamma)} > \frac{a_2 - b_2}{d_2 - c_2} \tag{76}$$

For $\alpha > \delta > \beta, \gamma$, the ratio $\frac{m(\alpha)-m(\gamma)}{m(\delta)-m(\beta)}$ is increasing in the convexity of m . This follows from the inequality below:

$$\frac{m(\alpha) - m(\gamma)}{m(\delta) - m(\beta)} > \frac{m(\alpha) - m(\delta)}{m(\delta) - m(\beta)} \tag{77}$$

which is increasing in the convexity of m . Similarly, the ratio $\frac{m(\alpha)-m(\beta)}{m(\delta)-m(\gamma)}$ is increasing in the convexity of m . Hence, there exists a sufficiently convex warm-glow function m (or m with a sufficiently large slope between α and δ) for which equations (73),(74),(75) and (76) hold. Further, as the interval

$$\left(\max \left\{ \frac{d_1 + m(\delta) - b_1 - m(\beta)}{a_1 + m(\alpha) - b_1 - m(\beta) - c_1 - m(\gamma) + d_1 + m(\delta)}, \frac{d_2 + m(\delta) - c_2 - m(\gamma)}{a_2 + m(\alpha) - b_2 - m(\beta) - c_2 - m(\gamma) + d_2 + m(\delta)} \right\}, \min \left\{ \frac{d_1 - b_1}{a_1 - b_1 - c_1 + d_1}, \frac{d_2 - c_2}{a_2 - b_2 - c_2 + d_2}, \frac{\delta - \frac{\beta+\gamma}{2}}{(\alpha + \delta) - (\beta + \gamma)} \right\} \right) \tag{78}$$

is non-empty, there exists a μ_0 in this interval such that, σ is a Bayes Nash equilibrium and the bystander’s utility is decreasing as the proportion of morality increases from μ_0 . Following similar arguments it can be easily shown that the Samaritan’s Curse can occur under Scenario 1 when the warm-glow function m is sufficiently convex. This completes the proof. □

Sambit Mohanty holds a PhD in game theory from the Industrial Engineering and Operations Research Department at Indian Institute of Technology, Bombay. His research studies the conflicting effects of morality in heterogeneous societies.

Jaideep Roy is a Professor of Economics at the University of Bath. He focuses on political economy and economic theory, and has published widely in journals such as *Games and Economic Behavior*, *American Economic Review*, *International Economic Review* and the *Journal of Development Economics*.

Cite this article: Mohanty S and Roy J. Individual versus group morality: the role of information. *Economics and Philosophy*. <https://doi.org/10.1017/S0266267124000178>