August 4, 2021

Convention, Morals and Strategy: Greta's Dilemma and the Incarceration Game

Kaushik Basu

Department of Economic and SC Johnson College of Business
Cornell University
Uris Hall
Tower Road
Ithaca, New York 14853

Email: kb40@cornell.edu

Abstract

Conventions and leaders are believed to be the two pillars of justice and order in society. This paper evaluates this proposition and draws attention to two intriguing ways in which these pillars can malfunction. The argument is constructed by creating two new games, Greta's Dilemma and the Incarceration Game. An awareness of these problems can help us use our 'moral intention' to reexamine our own collective behavior and to design prior conventions, which limit the power of the leader.

Key words: Just society, Conventions, Morals, Greta's Dilemma, Incarceration Game

Acknowledgements: For many specific ideas discussed in this paper, I have, over the last several months, received helpful comments and suggestions from several individuals, including Alaka Basu, Vidya Atal, Larry Blume, Devajyoti Ghose, Matthew Braham, Joe Halpern, Ajit Mishra, Peter Katzenstein and Derk Pereboom. I am grateful to them all.

Convention, Morals and Strategy:
Greta's Dilemma and the Incarceration Game

1. Introduction

We understand selfishness better than unselfishness. This is so for a reason. Adam Smith's seminal work (Smith, 1776) and discovery that, in many contexts, the order that we witness prevailing in the economy and society can be explained by the self-interest of the individuals constituting the society was so surprising that it triggered a big reaction. There were the ideologues who used this to revel in their selfishness, rationalizing that for order and justice in society they need not tire their hands; Adam Smith's 'invisible hand' will do that for them. There were also the legions of moral philosophers, political economists and even mathematicians who treated Smith's hypothesis as a challenge to formalize and analyze the extent to which the hypothesis was valid and could be relied upon when thinking of national policy. Contemporary economists owe a huge debt to this response to Smith. I am referring not to the celebration of selfishness, but the outpouring of research. From Leon Walras (1874) to Ken Arrow and Gerard Debreu (1954), the research that went into this field, bringing modern techniques of general equilibrium to analyze these old philosophical ideas, put us on a sound footing.

The understanding of behavior that goes beyond self-interest to matters of justice, fairness, and morals, has trailed the above research. The reason is that, while the early ideas go back to the seventeenth and eighteenth century, to Thomas Hobbes and David Hume, and arguably even earlier, to Plato and Aristotle, the formal tools we need for this analysis were not available till more recent times. I am referring to game theory, such as in the works of von Neumann and Morgenstern (1944) and Nash (1951). The use of game theory to shed light on these fundamental building blocks of a successful society is more recent. From a few contributions in the mid-twentieth century (such as Braithwaite, 1955; Runciman and Sen, 1965), there is now a gradual build-up of research that is beginning to reexamine these ancient themes, and creating hope for new theories of justice and order that are founded on a formal structure which we can actually put to test in laboratories and in the field.

This is what makes the new book by Peter Vanderschraaf, **Strategic Justice: Convention and Problems of Balancing Divergent Interests** (2019) so important. It spans the breadth of this subject, drawing on tools of modern game theory to give formal shape to early ideas which are important but hard to critique or to use in practical contexts because of the blurriness of their contours. This is a book that should be treated as important for providing building blocks for this field of enquiry, rather than for being a curtain call. The present paper joins the discourse in that spirit. My aim here is not to course over the large canvas of this book but to bring into sharp focus some specific themes discussed in it, by using some concepts and results from game theory and economics to bear on them. My own sympathy for the Humean agenda and the more critical engagement with Hobbes will be evident in what follows. In the process of this engagement this paper presents two new results, one of them concerning group moral responsibility and the other

raising some paradoxical questions about the role of knowledge and knowledge hierarchies in empowering the leader.

One conclusion of this analysis is that is there is no magic bullet for justice and the structuring of a fair society. This is not just a case of human beings not having *found* the magic bullet, but of the magic bullet not *being* there. Nevertheless, the research and its pursuit is important. It enhances our understanding and equips us better to solve the problems as and when they occur.

2. Conferred Morality and the Moral Intention

Conventions can help us select a better outcome from among numerous equilibrium outcomes available to society, without having to appeal to anything beyond each individual's interest. This is true of the classic repeated Prisoner's Dilemma, and Farmer's Dilemma (Vanderschraaf, 2019). But this only happens when these games are played with no commonly-known terminal date.

In such games, typically, there are numerous equilibria and it is easy for players to have a mismatch of expectations and end up on a disastrous outcome. Vanderschraaf, citing Schelling (1960) and Lewis (1969), rightly points out the critical role that conventions play in such situations. They create consistent expectations around a focal point or outcome in the minds of the players. Justice is, in this sense, a convention, that helps an individual form beliefs about the beliefs and behavior of another individual, and to have a sense of what the other person will do in response to what one does now[1]. These beliefs can help sustain an optimal outcome for society.

There are two caveats to this. First, 'optimal' in these contexts refers to Pareto optimality, that is, an outcome, where it is not possible to make one person better off without making someone else worse off. As is well known, Pareto optimal outcomes can be obnoxious. An oppressive society that exploits one group of people mercilessly, leaving no room for further exploitation may well be Pareto optimal. To prevent such oppressive 'optimal' outcomes we may need to go beyond self-interest.

Second, even when a genuinely good outcome is possible via a self-enforcing convention for a repeated game, this solution breaks down as soon we have a terminal date in games like the Prisoner's Dilemma or the Farmer's Dilemma, a last move by one player or both players, because of the well-known backward-induction argument.

[1] It is possible to take this logic further, to the rule of law, and argue that, in the very end, the law is *nothing but* a focal point, a mechanism for building a consistent set of beliefs among the citizenry (Basu, 2018). Of course, punishment plays a role. I respect the speed limit law for fear of punishment, but the traffic police punish me for violating the law because of their fear of punishment from the boss of the department, should they fail to do their job. In the end, it is an interlocking system of beliefs that explains everybody's behavior and the adherence to the law.

In such situations, once again, there is no escape from recognizing the need for checks on one's self-interested behavior (Vanderschraaf, 2021), either by norms, which may be like programmed responses in the human head, or by a conscious commitment to some principle of fairness or justice—essentially a 'moral intention', based possibly but not necessarily on empathy, which is an innate human trait (admittedly with variations across individuals and societies).

What I want to do now is to demonstrate that such a solution which brings in morals and tries to steer society to a genuinely good outcome can also break down in a rather interesting and almost paradoxical way, thereby leaving us with an open philosophical question. What is an ethically right action is a subject of enormous writing and debate. I use 'being moral' as a form of act consequentialism, where one puts some weight to the well-being of others. The formal method used here is what Vanderschraaf (2019, p. 150) suggests when he talks of agents who "get some satisfaction from the shares their partners receive as well as their own shares."[2]

It is important to understand that the paradox I am about to describe, referred to as 'Greta's Dilemma', happens only in strategic, game-theoretic environments, the kind that Vanderschraaf's book is concerned with, and not in a traditional, rational-choice setting. This is the reason why our intuition, founded in traditional settings, is jarred by the outcome in Greta's Dilemma. Hence, before I get to this new game, I want to quickly illustrate what 'being moral', in the above sense, does in a traditional setting.

Suppose an individual, call him person 1, has to choose one action from a set X of available actions. Each action $x \in X$ is denoted by a pair, (x_1, x_2), where x_1 denotes the payoff earned by individual 1, that is, the person making the choice, and x_2 is the payoff earned by the bystander. The bystander may be the future generation—henceforth, she. She cannot influence person 1's choice but is affected by it. Let us denote the choice that person 1 makes if he is completely selfish by $x^* \in X$. Then, by definition, $x_1^* \geq x_1$, for all $x \in X$.

Now suppose person 1 becomes moral, in the sense that he treats each unit of payoff earned by 2 as equivalent to δ units earned by himself, where $1 \geq \delta > 0$. Let $z^* \in X$, be the choice he makes once he becomes moral. It is obvious that in this kind of a non-strategic environment, the person making the choice becoming moral is always good for the bystander. That is, $z_2^* \geq x_2^*$.

To see this, suppose that the above inequality does not hold. That is,

$$x_2^* > z_2^* \qquad (1)$$

The fact that person 1, when selfish, chose x^*, implies $x_1^* \geq z_1^*$. But this, coupled with (1), implies

[2] In this paper, when I talk of being moral, I know I am stepping into controversial waters about what gives actions moral worth (see Markovits, 2010). If I help a poor person because I get joy, some would argue that this act is not moral. This controversy however has no bearing on what this paper is concerned with. I am *defining* moral behavior as one where a player maximizes not just her payoff but also gives weight to the payoffs earned by the poor and the marginalized. I do often use the word empathy but that is purely for linguistic ease. The reader is free to impute the reason for such behavior. My result will be unchanged.

$$x_1^* + \delta x_2^* > z_1^* + \delta z_2^*.$$

But this is impossible since, when he is moral, he chooses z^*. The contradiction proves that the person making the choice becoming moral is always desirable (or, more correctly, is more desirable or the same) for the bystander or the future generation. This is the reason why all kind and environmentally-conscious people, like Greta Thunberg, urge us to have empathy for the future generations.

Keeping this as backdrop, note that group moral responsibility and the moral obligation of groups is a large area of discourse (see, for instance, Feinberg, 1968; Sartorio, 2004; List and Pettit, 2011; Schwenkenbecher, 2019). I have engaged with this elsewhere (Basu, 2021a) and shall not here delve into the various controversies.

My argument here relates to another literature that recognizes that most societies are morally diverse. In fact, moral diversity is a defining feature of modern societies (Thrasher and Vallier, 2015; Moehler, 2018)[3]. This is in contrast to much of mainstream economics, where individuals are treated as largely homogenous, in the sense of everybody being amoral. Players may have diverse *preferences* but they are not other-regarding. In Basu (2021a) I consider the case of a fully moral society and, separately, a fully amoral society. What happens if there is diversity, in the sense of some individuals being moral and some amoral? The game, Greta's Dilemma, shows the paradoxical implication of this diversity being achieved by some individuals becoming moral creatures.

How these debates are resolved can make important difference to how we draft laws and how we choose to behave. When it comes to self-interest, thanks to game theory, we have reached much clarity on attributions of group behavior to individual choices. Seeing a group of villagers over-grazing the commons and hurting their own well-being, we do not lament that the villagers are not acting in their self-interest. We realize that it is in fact their self-interested behavior that is the source of the problem. The Prisoner's Dilemma and the Traveler's Dilemma are illustrations of this, as is the Farmer's Dilemma of Hume (1740). They can help us design policies in the real world for solving the commons problem and crafting agreements among nations to tackle global climate change. We are today aware of many of the pitfalls of individual rationality for collective well-being. The same is not true for moral behavior.

In Basu (2021a), I develop the game, the Samaritan's Curse, to show that an entire society turning moral can make the society behave immorally. In some sense, this is a flip-over of the Prisoner's

[3] It is worth pointing out here that the result obtained in Basu (2021a), of a scenario where everybody turning moral results in a morally inferior group outcome, does rely on moral diversity. This would not be a surprise if the moral conflict was deep, or to use Moehler's (2019, p. 3295) words, "in deeply morally diverse societies". In my paper, as well as here, the morals are much more common. There is no "foundational disagreement" (Quong, 2005) among the citizens. Morality implies being concerned about the deprived, or the bystander, like future generations that cannot influence us directly but have to rely on our actions. Individuals have their self-interest but this is tempered by this shared 'moral concern.'

Dilemma to a moral domain. However, a single individual's turn to morality in the Samaritan's Curse causes no reversal in group morality. The game developed in this section, Greta's Dilemma, shows that a tiny infusion of morality, on the part of even one person in a society, can cause a paradoxical reversal. This provides an interesting twist to a theme that has been investigated before. Sartorio (2004) argues how it is possible for a person to be responsible for an outcome that was not caused by any action taken by the person. What I show here is that, in some situations, it is not so much the person's action as who the person is that causes the group to behave immorally. And to make matters worse, it is this single person *becoming moral* (in the sense of Vanderschraaf, 2019) that results in the entire group behaving immorally.

Consider the interaction between two individuals, Johnny Won and Jaya Tu, whose names, luckily, can be shortened without phonetic distortion to 1 and 2. They will be called 'players', since these are people who choose between alternative actions. In this 'game', player 1 chooses between actions, A and B, and 2 chooses between actions, C and D. Hence, the 'outcomes' of the game are (A, C), (A, D), (B, C), and (B, D), where, in each pair, the first symbol denotes Won's choice and the second symbol denotes Tu's choice. The earnings of the players are shown in the payoff matrix labelled "the Basic Game", on the left, in Table 1, where player 1 chooses between the rows and 2 between the columns. The payoffs in each cell are, respectively, 1's payoff and 2's payoff. Thus, if 1 and 2 choose A and C, respectively, they earn 100 and 101. And, in game-theoretic language, if they choose (B, C), they earn (101, 100). For simplicity, we may think of the payoffs as dollar earnings.

To give a realistic story to the symbols, think of A as the environmentally-friendly activity of 'Agriculture' and B as the more damaging business of operating 'Brick kilns'. And C stands for 'Coal-mining, which can do a lot of damage to the environment, and D for the greener activity of 'Dairy farming.' The significance of these will be obvious when I consider the fall out of these choices on the bystander, in this case the future generation.

Given individuals' wish to maximize their own earnings, a 'Nash equilibrium', henceforth referred to simply as 'equilibrium', is an action pair from which neither player has an interest in deviating unilaterally. It is easy to see that in this game the only equilibrium is (B, D), where they earn (101, 101). If 1 deviates to A, he earns 100 and if 2 deviates to C, she earns 100. So neither will deviate.

Now suppose this society has a 'future generation' that is poor (thanks to the damage to the environment done by the present generation). The future generation's well-being is entirely dependent on what the current generation does. If 1 and 2 choose the environment-friendly actions A and D, the future generation gets 8. If they choose the environmentally worst option (B, C), the future generation gets 0. If they choose the environmentally more mixed options (B, D) or (A, C), the future generation gets payoffs of 4 or 2. This is summed up in the future generation's Earnings described in the right-hand matrix in Table 1. Henceforth, the collection of the two players' payoffs (in the left-hand matrix) and the future generation's or the helpless bystander's payoffs (in the right-hand matrix) will be referred to as a 'society'.

Let me begin with standard game theory and neoclassical economics, in which the future generation's payoff is of no consideration to how players behave. They are solely interested in their own payoffs and this leads to the equilibrium (B, D), where the two players get 101 each and the bystander gets a miserable 4.

To most observers the behavior of the group of players will appear unethical. The current generation is well-off, each person earning $100 or more. Moreover, how they choose can make only a $1 difference to their income. The future generation's condition is likely to be much more precarious, with much lower payoffs. Note that if the players moved from (B, D) to (A, D), the future generation or the bystander would get $8 instead of $4. It appears totally unethical not to make this small sacrifice.

Table 1. A Society

The Basic Game

	C	D
A	100, 101	100, 100
B	101, 100	101, 101

Future Generation's Earnings

	C	D
A	2	8
B	0	4

Now suppose, player 1 has a chance meeting with Greta Thunberg and learns about the kind of ethics mentioned above, the importance of being concerned about our progenies especially because they do not have a voice in our current deliberations. Won learns that a minimal requirement of morality is to be mindful of how one's actions affect poor bystanders. He takes Greta's advice to heart, as indeed all of us should, and becomes a human being, who takes the welfare of future generations into consideration in making his decisions. For simplicity assume, player 1 now gives the same weight to future generation's earnings as he does to his own, in the spirit of Vanderschraaf's idea of empathy.

This causes the game to change. There is now one selfish player (player 2) and a moral one (player 1). What the game converts to is illustrated in Table 2. I call this game 'Greta's Dilemma'. This new game is created by adding the future generation's earnings to player 1's earnings, since that is what player 1 now tries to maximize.

What is the outcome of Greta's Dilemma? It is easy to see, it has only one equilibrium, (A, C). The old outcome, (B, D), is no longer an equilibrium since if player 2 chooses D, player 1 will choose A *in order to help the future generation*. And knowing that player 1 would choose A, player 2

would choose C. Check that no other pair of choices is stable. We invariably end up at (A, C), thanks to player 1's conversion from an amoral agent to a moral person[4].

What happens to the future generation? Before Won met Greta, the future generation earned 4. After player 1 comes under her influence and becomes an environmentally-conscious person, determined to help the future generation, the future generation's well-being ends up dropping, to a shockingly low payoff of 2.

Table 2. Greta's Dilemma

	C	D
A	102, 101	108, 100
B	101, 100	105, 101

Outside observers, seeing the atrocious behavior of this group of two rich persons pushing the precarious future generation to a low well-being of 2, will find it difficult to understand that this is happening *because* one of them has become moral, with a commitment to help the future generation. Clearly, morality on the part of an individual, *in a strategic environment*, does not necessarily translate into a moral outcome.

To understand Greta's dilemma, let us call an outcome in which anybody gets a payoff less than \$3 a 'bad outcome'. Thus (A, C) and (B, C) are bad. It is easy to see, if a bad outcome occurs, Won has no responsibility for this, since the outcome would be bad no matter what he chose. But Tu has moral responsibility. If she does D, the bad outcome can be averted for sure.[5] The problem arises from trying to reconcile this with the fact that it is player 1's visit to the seminary and *becoming moral* that causes the bad outcome.

One possible way of countering this is to argue that player 1 can surely see that it is his being moral that causes the harm and so it is better to *act* amoral and be selfish. Then, the game will end up at (B, C), where the bystander does better, by getting 4. But this is akin to arguing that, confronted with the Prisoner's Dilemma, players will have the sense to behave as though they

[4] While this specific game is illustrative, it is arguable that this is real problem which has implications for how we shape policy in the real world (Basu, 2021c). It should be possible to do some interesting laboratory tests with Greta's Dilemma to see if the predicted moral setback is likely in reality. The test should take the form of making players play the basic game in Table 1, without letting them know of the effect their play has on the bystander. Thereafter, tell player 1 about the bystander and maybe even do some priming about moral responsibility, and then make them play the same game, to see if player 1 flips from choosing B to A. There has been work concerning the manipulation of beliefs and focal points in laboratory settings (Dasgupta and Radoniqui, 2021), which can suggest ways of imputing morals and values into actions.

[5] In the context of game theory and morality, similar ideas are discussed in Braham and van Hees (2012).

are not selfish and achieve the good outcome. Such an argument is dismissed by game theorists on the ground that it always pays to deviate after persuading the other player that you are someone else.

The analogue of this in Greta's Dilemma is for 1 to pretend to be amoral, drive the game to (B, D), and then make a last-moment switch to A and achieve (A, D) and help the bystander get 8. To this the game theorist's rebuttal will be (and philosophers should surely agree): player 2 will expect this and so play C in anticipation and we are back to (A, C).

There is however an interesting difference between the selfish I, and the moral I that traditional game theory is unable to grapple with. This lies more in the domain of philosophy. In game theory, we treat self-interest (that is, the urge to maximize one's own payoff) as innate. But it is arguable that being moral entails an element of volition. So it is possible to argue that individuals can choose to be moral (or not), after examining the context, in this case, the game. If they are sufficiently far-sighted, they may decide, in certain contexts, not to *behave* morally.[6]

Greta's Dilemma also takes us beyond moral philosophy to connectionism in the cognitive sciences, which recognizes how one human mind consists of large networks of neurons, each neuron oblivious of its role. There is now an effort to bring this to the social sciences and recognize there may be large networks of humans behaving in ways that each person is unaware of (Clark and Chalmers, 1998; Fioretti and Policarpi, 2020). It is critical to study such organism in its entirety. But this will, in many situations, make it futile to try to morally evaluate the individuals that comprise the collectivity.

I would argue that, for now, hope has to lie in the kind of suggestion put forward by Putnam (2005, p.24), drawing on the work of Emmanuel Levinas: "For Levinas, the irreducible foundation of ethics is *my* immediate recognition, when confronted with a suffering fellow human being, that *I* have an obligation to do something. [Even if I cannot actually help,] not to feel the obligation to help the sufferer at all, not to recognize that if I can, I must help … is not to be ethical." Note that Putnam is not contradicting the dictum 'ought implies can'. He is not asserting you ought to help someone whom you cannot help, but you must feel the obligation to help.

Building on this, it may be argued that, even if a good outcome is beyond current reach, we must nurture and keep alive what may be called the 'moral intention,' which is the intention to achieve the good eventually. If the moral *intention* is deadened, and we end up in a morally-bad outcome, we would treat this as a *fait accompli* and do nothing. It is our moral intention that makes us want to step beyond the game under consideration, and think of how we may alter our behavior, such as by adopting deontological ethics or altering the rules of the game by imposing taxes and fines on players or to act selfish even if one is not so, in contexts like the Greta's Dilemma game.

[6] A related problem has been discussed in the coalition-formation literature (Aumann and Myerson, 1988; Genicot and Ray, 2003; Ray and Vohra, 2015). But the choice of *morals* raises conceptually distinct matters.

The upshot is: there is no easy getting away from the core message of Greta's Dilemma. When we see a group behaving badly we must not assume that the outcome reflects what the individuals in the group desire. Most people do not allow the thought that a group of leaders may not want to do what leaders as a collective often end up doing. They may be in a trap the same way that Havel (1986) conjectured that the leaders of a post-totalitarian state may have no exit route. This is indeed a dilemma that Greta Thunberg, with her good moral intention, has to be aware of. Individuals being concerned about future generations not only may not aid help those generations, but may actually end up hurting them. Strategic environments, the heart of game theory and what reality is mostly about, throws up unexpected challenges.

Let me close this section with two footnotes, too important to be relegated to the bottom of the page. First, I have commented on the need for prior conventions to protect us from getting trapped in a bad equilibrium; and this will crop up again below. In much of the literature on games and justice, the focus is on a *point* outcome of the game, typically a Nash equilibrium, which is salient, with the salience achieved by virtue of the convention. The convention targets a particular (presumably desirable) outcome. There is, however, a need to go beyond this and consider set-valued equilibria, that is, conventions which take the form of my agreeing not to step outside a particular set of actions, your committing not to go outside a set of actions and so on. Especially the idea of a constitution is best conceived of as a Cartesian product of subsets of each person's available set of strategies. After all, that is what a convention in reality does. It prohibits the use of certain actions; it does not usually pin you down to a specific action. There is no need to develop this idea for the purposes of this paper but this is an important lacuna in the literature, and deserves to be pursued in the future.

Secondly, one word concerning the need to distinguish immorality and the labeling of immorality. The above analysis was concerned with making us more circumspect in holding individuals morally responsible for collective immorality. However, there is a case for distinguishing between holding people morally responsible and *saying* that you hold them morally responsible (Basu, 2021a). It is known that to be criticized for being immoral is not something people like. It is a kind of punishment. In other words, if people are described as immoral for making certain choices, this may change their behavior so that the group makes better choices. This means that there may be a consequentialist reason to describe people as immoral for certain kinds of behavior even when we do not really hold them morally responsible.

3. A Humean View of Leadership

If society has a propensity to end up in an outcome that is anarchic or one that is morally regrettable, as happens in Greta's Dilemma, what is the way out? We characteristically fall back on the idea of a prior convention or having a leader, who can steer society out of such an outcome. This is where the line taken by Hume is often treated as a breakaway from much of earlier thinking, and in particular, that of the analysis of Hobbes (Vanderschraaf, 2019, Chapter 6). Efforts to reconcile these two streams of thought are few and far between (see Gauthier, 1990; Sugden, 1986). I have recently taken a rather different line, best described as a Humean

solution to the conflict between Hobbes and Hume (Basu, 2021b). In some ways, the solution offered is close to the position taken by Vanderschraaf (2019). I want to briefly recount this by way of prelude to presenting a new idea in the next section.

At times, it appears that people need top down control to behave in certain desirable ways. For instance, to ensure that people respect the speed limit law, we need traffic police to keep vigil and punish errant drivers. But once we take the next step and ask why the police will do the job they are expected to do, we end up with the same kind of answer. Someone will punish the police who does not punish the driver who violates the speed limit law. In the end, if there is order in society, it must be caused by nothing but a web of beliefs of what each person expects from others. This was the stunning insight of Hume. In the language of modern game theory, a stable social outcome is nothing but a focal point of the game of life. A convention does nothing but create a focal point (Lewis, 1969); the law does nothing but create a focal point (Basu, 2018).

At first sight it appears that, according to this Humean view of order, there is no need for a leader. This is not true for an interesting reason. On a little thought, it becomes clear that there are situations in life where players do not know in advance what game they will be playing. This is most obvious in a war or a revolution. The army may be confronted by a sudden situation, where every soldier has to decide, which of the 10 nearby peaks of hills, they should run up on. If they all climb the same peak, they will win the war. Otherwise the enemy will vanquish them. Each player may have his or her own favorite peak but all are agreed that between being on the same peak and different ones, they are all better off being on the same.

Since the game appears before them suddenly (as happens in war and times of conflict), there will be no time to discuss and agree on a focal point or a convention, even a limited one (in this case which of the 10 peaks they should all climb). The way to solve this is to have a focal *person*, chosen in advance. This extension of Schelling's idea of a focal *point* to a focal *person* can solve the problem. The trick is to have one person designated commander of the battalion. After the game suddenly emerges there may be a cacophony of voices pointing to different hill tops, but everybody knows everybody will listen to the voice of the commander—the focal person[7]. There is no surprise that armies have well-defined leadership.

This shows that, while we do need leaders in some situations, the power of the leader is, in the end, rooted in the same web of beliefs that Hume was referring to. The leader is powerful because each soldier knows that the other soldier will listen to the leader and each soldier knows that the other soldier knows that others will listen to the leader, and so on. The power of the leader is not an exogenous force but a product of the same lateral beliefs that support a focal point. It is interesting to see a grudging recognition by Hobbes in Chapter XIII of **Leviathan** that the leader's power does not come from the leader's strength: "[W]hen all is reckoned together,

[7] In Basu (2018), based on an argument of Myerson (2017), I argue the need to extend Schelling's concept of the focal *point* to a focal *person* or player. The idea of a leader who can create focal points in the action space emerges out of this, as I formally show in Basu (2021b), though the basic conceptual idea goes back to David Hume.

the difference between man, and man, is not so considerable... For as to the strength of body the weakest has strength enough to kill the strongest".

Once this view of the leader is formalized it becomes clear that one can get some deep insights into the impact that leaders can have on society. We see that, contrary to Hobbes's claim, there are important contexts where having multiple leaders can benefit the people by each leader acting as a curb on the others' power. We also find that there are situations where the citizenry would be better off with no leader than the one Leviathan (Basu, 2021b). In such situations, in case there is a leader, there is a genuine democratic need to have the leader removed from the position of power.

What I want to do here is go a step further and show how leaders can create mechanisms of oppression that block avenues whereby they can be evicted from power. I demonstrate this formally by constructing a model of the **Incarceration Game**. This results in stability but the stability of oppression. How can such oppressively stable outcomes be blocked? This brings us back to the theme of Vanderschraaf (2019). We need to have in place a convention, before the leader comes to the seat of power, a convention that blocks in advance the possibility of the kind of action the evil leader might take.

This also connects us to a classical debate inspired by Tullock's (1971) "paradox of revolution" that Vanderschraaf (2008) had joined. Tullock's argument was concerned with the public goods problem of revolution, whereby even if every citizen wants to overthrow "a vicious, corrupt, oppressive, and inefficient government" (Tullock, 1971, p. 89), they may not be able to because each person's joining the protest makes little difference to the outcome, and the risk to each person is likely to be too high. So no one joins and the revolution, which would have been successful if everybody joined, does not happen. Vanderschraaf dislodged Tullock's argument by an innovative combination of game theory and tipping point theory.

In the section that follows, I take this line of argument further but by a novel route. What this does is to also highlight how evil leaders can squash revolts by the clever use of knowledge hierarchies among players in a game. The aim of my model is to provoke research into foiling the strategy of the evil Leviathan

4. Rebellion and Reprisal: The Incarceration Game

In Chapter 6 of his book, Vandershraaf discusses the risk of malevolence on the part of the Leviathan. If all citizens live by the law but the Leviathan is beyond the reach of the law, is there not a need to have a strategy to tame the Leviathan? It is interesting to find Vanderschraaf (p. 215) marshalling Hume to argue the need to keep the ultimate authority "somewhat in check by the threat of revolt." It is this same concern that reminds us of the need for a convention or a constitution, *prior to the arrival of the leader* to the helm of power, thereby setting limits on what the leader can do. The Leviathan may need taming too. The analysis in the book can be taken further to gain some deep insights into the problem. I present here a rather intriguing

perspective, which dictators have (I suspect) known and used long before the birth of game theory. But we can understand this better by giving it formal, game-theoretic shape.

The world has seen revolts and rebellion since ancient times. To understand these at an abstract level is as fascinating as trying to understand the order that prevails in society, whether because of the invisible hand of the market or the iron hand of the Leviathan. However, it seems easy to throw up our hands concerning revolt, since by their very nature, and in contrast to order and harmony, revolt is difficult to model. But we have made big strides in understanding anarchy and the state of nature (Vanderschraaf, 2006; Moehler, 2009), and so there is hope for revolt too. In this section I take on a thin slice of this big agenda.

History is full of examples of revolution breaking out amidst seemingly peaceful times, and rebellion being silenced and having its back broken on the verge of success. Even within the last few decades we can find examples of both. From the Tiananmen Square massacre or 'the June 4th Incident,' as the rebellion of 1989 is often called, through the Jasmine Revolution in Tunisia or the more widespread Arab Spring in the Middle East that began in 2010 and led to the collapse of several authoritarian leaders, to contemporary incidents like the protests against Lukashenko's brutal regime in Belarus, one sees a gathering steam but has also frequently seen a quietening down, with the streets abandoned by protestors because of the looming threat of reprisal. Something similar is going on in Myanmar, where after the coup of 1 February 2021, the military leaders who master-minded the coup have killed over 800 of the citizens protesting against the military dictatorship.

There can be subtle play between beliefs and strategy that can give us insights into both the successes and failures of rebellion. To understand this, consider a tyrant who has fallen out of favor with the citizenry raring for change.

It is not difficult to see why such widespread opposition to the government at times manifests in open revolt. For that to happen, the need is for a galvanizing incident, such as the self-immolation of Tunisia's Mohamed Bouazizi, a street vendor who set himself on fire on 17 December 2010, thereby creating a focal point, and sparked the Jasmine revolution. Or we may need a 'focal leader' who can help the citizenry coordinate its actions.

Digital technology has facilitated such coordination. No one wants to go out and protest alone; it is too dangerous because the authoritarian leader can have the person arrested, incarcerated or executed. But if thousands of people go out at the same time to protest, each of them is relatively safe, since there are limits to how many people a tyrant can arrest and hurt. The need therefore is to coordinate on the time and on the place, which is the classic focal point problem (Schelling, 1960; Sugden, 1995).

We saw the power of new technology in achieving this in Tunisia's Jasmine Revolution. People could exchange messages and make sure that they would not be caught protesting alone or in small numbers out protesting. This was made possible by an exile, Amira Yahyaoui, and by the availability of digital technology. In 2005, Yahyahoui, who was then a young political dissident,

was beaten up by the secret police of the Tunisian leader, Ben Ali, and sent into exile in France. From there she began to help coordinate dissidents using social media. In 2010 she organized an event via the web whereby Tunisian activists took to the street simultaneously. The protests gathered momentum and on January 14, 2011, Ben Ali fled the country, taking refuge in Saudi Arabia. As Boumlik and Schwartz (2016) noted, "Yahyaoui's influence within the collective movement contributed to regime change through her advocacy from her exile in France via social networks."

This was a happy ending but there are examples of rebellions that were snapped on the verge of happening. Belarus's opposition leader, Sviatlana Tsikhanuouskaya, has been in the relative safety of exile, in Lithuania and Poland, and has tried to galvanize opposition. This did build up for a while but the streets of Minsk have since been deserted by protestors.

This gives rise to fascinating questions about why some uprisings succeed and some fail; and what methods tyrants use to foil coordinated rebellions. The hope is that a better understanding of the modus operandi of oppressors will enable us to construct laws and conventions to foil such oppression.

I want to illustrate how tyrants can suppress dissidence by using strategic analysis, the hierarchy of knowledge and game theory. The game theory is so intuitive that one needs no prior knowledge of game theory to understand it. There is therefore no surprise that tyrants have actually used similar methods. We were lucky that some like Ben Ali did not figure this out.

To explain this with a parable consider a leader of an ancient city, who was once popular with the masses especially the young, has turned into a tyrant, incarcerating, torturing and killing anyone who opposes him and also playing the fiddle when the citizens suffered (any resemblance to the Fifth Emperor of Rome is entirely spurious). Suppose this ancient city has a population of 10 million plus one, that is, 10 million citizens plus the one leader. The fact that the population of this nation is virtually the same as that of Belarus is pure coincidence.

All citizens want to throw the tyrant out of power and let me assume that if one million or more people come out and join the protest, the leader will be deposed. Next assume that all citizens have agreed about the date and time of the protest. The focal point has been created. Each citizen has to choose between joining the protest (action P) or being silent (action S). Assume that, other things being the same, every citizen prefers to protest than be silent. Let each citizen's payoff from choosing P be 100, and the payoff from choosing S be 0. Whether a person will choose P or S depends on the expected punishment for protesting[8].

[8] For simplicity I shall assume that each person treats the success of the revolution to be independent of one's own action. In a large population this is not unreasonable, and it keeps our analysis simple. So the satisfaction of 100 from protesting is a pure joy of participating in this important task of helping the nation. This keeps us out of the public goods problem that concerned Tullock (1971), and is not germane to what I am doing now. It should be emphasized that my assumption is not unrealistic. People do get satisfaction from voting even when they know their individual vote is unlikely to matter; people do feel good not throwing plastic into the river even though one piece of plastic is unlikely to do any harm.

The leader wants to stop the revolution. So he starts to think what he can do. Let us suppose he has the ability to arrest and incarcerate protestors for life. There is however a limit to jail capacity. Only 10 persons can be jailed. So if more than 10 people go out to protest, he cannot arrest all protestors. Assume that the pain of going to jail is 200. Hence, if by protesting one faces a probability p of going to jail, then a person will protest if:

$$100 - 200p > 0,$$

In other words, a person will not protest if

$$p \geq ½ \qquad (2)$$

I am making the innocuous tie-breaking assumption that, when indifferent between P and S, one chooses S.

If all 10 million are ready to protest, it seems like a hopeless situation for the tyrant. If he announces he will arrest and jail 10 of them chosen randomly, the risk of being jailed is too small to deter anybody. All 10 million will protest and he will be thrown out of power.

He can stop some people by targeting them. Thus if he announces 10 names of people he will arrest if they are found protesting, none of them will choose to protest. But that still leaves many more who will protest and the rebellion will succeed. He can stop some more people from protesting by using the strategy of announcing up to 20 names, making it clear that if these people are found protesting, he will select 10 of them randomly and jail them. Clearly, given (2), above, none of those 20 will join the protest. But still that is not enough to stop the revolution which will go through as long as one million people are ready to protest.

The tyrant is reconciled to a revolution, knowing that his announcement of incarcerating 10 protestors cannot stop the protest. Feeling sad, he decides, now that his fate is sealed, he must do one good act before he is thrown out of power so that he is at least on the right side of the sky-god, Jupiter.

Suppose that this is a city where the age of each person is known (and assume for simplicity that all citizens have a unique age). The tyrant announces that while he will keep his word of jailing (for life) 10 of the protestors, he will do so with an element of charity. With an eye on pleasing Jupiter, he declares that, of all the people protesting, he will the 10 oldest, so that the jail time served by the citizens is minimized. Recall the imprisonment is for life. So if people have the same life expectancy, jailing the 10 oldest protestors will minimize aggregate jail time. He pauses the fiddle and announces his decision from the roof top, for all to hear.

He is reconciled that the protest will be widespread, and he will be deposed. His only hope is Jupiter will be kind to him. Then the day of the protest arrives. As he waits and watches, he is

caught by surprise. Not a soul comes out to protest, not even teenagers, who are nowhere near old. The rebellion is foiled. His tyranny can continue unabated.

It is not hard to see what happens. Let me call the game that the citizens play, after the public announcement by the leader, the **Incarceration Game**. For ease of exposition let me give citizens names. The oldest is called person 1, the second oldest person 2, and so on, the youngest person being person 10 million.

Note that, in the game of Incarceration, citizens 1 to 10 will immediately give up any thought of protest. They know that if they are out there, they will all be arrested and jailed. So each of them chooses option S. Since the individuals 11, 12, …, 20, can reason and get to know what the first 10 citizens will do, they will be aware that if they go out protesting, *they* will be jailed. So they will choose S. Now with citizens 1 to 20 choosing S, the same logic will make the next 10 choose S. This continues, with all 10 million people deciding to be silent instead of protesting.

The reasoning involved in the Incarceration Game is familiar territory in game theory and analytical philosophy. We have seen illustrations of this in the Surprise Test paradox (O'Conner, 1948; Quine, 1953; Scriven, 1953), the Centipede game (Rosenthal, 1981; Binmore, 1996) and the Traveler's Dilemma (Basu, 1994; Halpern and Pass, 2012; Rubinstein, 2016). It would be folly to treat these kinds of common knowledge reasoning as pure academic exercise[9]. Dictators do use this, often without full understanding of how it works, and they often succeed. We need to understand the role of this kind of strategic thinking and the power of layered knowledge.

In any real nation with a reasonable sized population, it is virtually impossible to make these kinds of rules of arrest, incarceration or execution common knowledge. But in the problem described above common knowledge is a sufficient condition; it is not necessary[10]. Rebellions can be foiled with less than that. Much depends in reality how close we can get to common knowledge. The top down oppression in Tunisia may not have done what is being done in Belarus today. A clever leader can devise ways of spreading the news of his strategy so that the layers of knowledge build up in ways sufficient to foil the protest. For instance, if in the above problem, it is difficult to take in the information of how each of the 10 million persons is ordered in the sequence of punishment, we could have the society broken up into 500,000 groups of 20 individuals, and specify which group will be jailed first, which group second and so on. The tyrant can make it clear that if there is a large number of people protesting, it is not ordinary citizens who will face the risk of arrest. The tyrant will arrest the opposition leaders. If the opposition leaders are not there protesting, he will arrest the journalists who criticize his regime. If journalists are quiet and not protesting, he will arrest some other category of people. Once there is a well-known order of

[9] The pervasive use of these kinds of knowledge hierarchies, from various academic disciplines, through British whodunits and real life situations, like eating pau bhaji on the beach are discussed in Sarangi (2020). It is also worth recalling that Rubinstein's (1989) Email game originated from the real-life problem of generals trying to coordinate an attack on enemy territory.

[10] If this problem was applied to a nation with a countably infinite population, common knowledge would be necessary. Anything short of infinite layers of knowledge would not work. In that sense, this would be akin to Rubinstein's Electronic Mail game (Rubinstein, 1989).

arrests, who they will come for first, who second, and so on, the backward induction argument sets in and no one will protest. The oppression of the tyrant will continue unabated.

There are of course complications in reality. There are individuals, such as Mahatma Gandhi, Vaclav Havel, Nelson Mandela, Martin Luther King, whose moral commitment is so deep as to make them 'irrational', in the sense in which we use the term in game theory. It may not be possible to get such persons to modify their behavior through threats. Often, change comes from the presence of such people. As Bernard Shaw put it more colorfully, in **Man and Superman**, "The reasonable man adapts himself to the world: the unreasonable one persists in trying to adapt the world to himself. Therefore, all progress depends on the unreasonable man."

Having a moral compass prompting one's behavior can be valuable for society. However, as Greta's Dilemma highlighted, there is no sure-fire solution. Good people, in strategic environments, can end up doing harm they never intended to.

My aim here is not to solve the problem—indeed, I am not able to—but to show how close to reality, the philosophical debates, that works such Vanderschraaf's and the growing literature on moral philosophy and game theory epitomize, come to. The task, in reality, is hard. Just as it is not easy for dictators to foil rebellion because they cannot design the punishment scheme and publicize them well enough for it to be common knowledge or near common knowledge or fail to deal with the few irrationally moral persons in society, we may not succeed in creating prior agreements and constitutions to restrict the power of the leader. The Incarceration game lays out the kind of challenge we have to meet in creating conventions to limit the power of the Leviathan.

5. Epilogue

The Incarceration game is meant to illustrate a route to political oppression that has been used, in some form or the other, knowingly or unwittingly, by tyrants. The formalization is done to help us, the citizenry, understand the grammar of oppression. It is true that I do not have a solution about how to block such behavior, excepting to point out that this brings us back to the topic with which the paper began, namely, the need for societies to have some form of convention in place before a leader comes to power. It also suggests the need for us to step beyond the game of life and think of advance collective action to steer the outcome to the just, fair and equitable one. The aim of the last section was to set the stage for such an inquiry.

A final word about a seeming point of conflict between the analysis of game theorists and philosophers. This pertains to the 'game of life', the total game that we human beings are engaged in (Binmore, 1995). We begin our analysis often by describing the game of life. Then when the outcome turns out to be a dismal one, we discuss how we need to correct it. Should be change our thinking from the standard game theorist's "I based" thinking to the collective one? (Gauthier, 1987; List and Pettit, 2011; Schwenkenbecher, 2019) Should we decide to abandon act-consequentialism in favor some deontological commitments? The trouble with these

questions is that they take us to beyond the game of life, which is, by assumption, a non-existent terrain. This is a virtual contradiction, the resolution of which must be treated as priority for making further progress in the area.

References

Arrow, K. and Debreu, G. (1954), 'Existence of an Equilibrium for a Competitive Economy,' **Econometrica**, vol. 22.

Aumann, R. and Myerson, R. (1988), 'Endogenous Formation of Links between Players and of Coalitions,' in A. Roth (ed.) **The Shapley Value: Essays in Honor of Lloyd Shapley**, Cambridge University Press, Cambridge.

Basu, K. (1994), 'The Traveler's Dilemma: Paradoxes of Rationality in Game Theory' **American Economic Review, Papers and Proceedings**, vol. 71.

Basu, K. (2018), **The Republic of Beliefs: A New Approach to Law and Economics**, Princeton University Press, Princeton.

Basu, K. (2021a), 'The Samaritan's Curse: Moral Individuals and Immoral Groups,' **Economics and Philosophy**, vol. 37.

Basu, K. (2021b), 'Why Have Leaders at All? Hume and Hobbes, with a Dash of Nash', **Homo Oeconomicus**, forthcoming.

Basu, K. (2021c), 'The Ground Beneath Our Feet,' **Oxford Review of Economic Policy**, forthcoming.

Binmore, K. (1995), 'The Game of Life,' **Journal of Institutional and Theoretical Economics**, vol. 151, No. 1.

Binmore (1996), 'A Note on Backward Induction,' **Games and Economic Behavior**, vol. 17.

Boumlik, H. and Schwartz, J. (2016), 'Conscientization and Third Space: A Case Study of Tunisian Activism,' **Adult Education Quarterly**, vol. 66.

Braham, M. and van Hees, M. (2012), 'An Anatomy of Moral Responsibility,' **Mind**, vol. 121.

Braithwaite. R. (1955), **Theory of Games as a Tool for the Moral Philosopher**, Cambridge University Press, Cambridge.

Dasgupta, U. and Radoniqui, F. (2021), 'The Republic of Beliefs: An Experimental Investigation,' IZA Discussion Paper, No. 14130.

Feinberg, J. (1968), 'Collective Responsibility,' **Journal of Philosophy,** vol. 65.

Fioretti, G and Policarpi, A. (2020), 'The Less Intelligent the Elements, the More Intelligent the Whole. Or Possibly Not?' mimeo: University of Bologna.

Genicot, G. and Ray, D. (2003), 'Group Formation in Risk-Sharing Arrangements,' **Review of Economic Studies**, vol. 70.

Gauthier, D. (1987), **Morals by Agreement**, Oxford University Press, Oxford.

Gauthier, D. (1990), **Moral Dealing: Contract, Ethics and Reason**, Cornell University Press, Ithaca, NY.

Halpern, J. and Pass, R. (2012), 'Iterated Regret Minimization: A New Solution Concept,' **Games and Economic Behavior**, vol. 74.

Havel, V. (1986), 'The Power of the Powerless,' in J. Vladislav (ed.), **Living in Truth**, Faber & Faber.

Hobbes, T. (1651), **Leviathan**, (Cambridge University Press, ed. R. Tuck, 1991).

Hume, D. (1740), **A Treatise of Human Nature**, (Oxford University Press, eds. D. F. Norton and M. J. Norton, 2000).

List, C. and Pettit, P. (2011), **Group Agency: The Possibility, Design, and the Status of Corporate Agents**, Oxford University Press, Oxford.

Markovits, J. (2010), 'Acting for the Right Reasons,' **Philosophical Studies**, vol. 119.
Moehler, M. (2009), 'Why Hobbes' State of Nature is Best Modeled as an Assurance Game,' **Utilitas**, vol. 21.
Moehler, M. (2018), 'Diversity, Stability and Social Contract Theory,' **Philosophical Studies**, vol. 176.
Myerson, R. (2021), 'Village Communities and Global Development,' **Horizons: Journal of International Relations and Sustainable Development**, vol. 18, 228-241.
Nash, J. (1951), 'Non-cooperative Games,' **Annals of Mathematics**, vol. 54.
O'Connor, D. J. (1948), 'Pragmatic paradoxes,' **Mind**, vol. 57.
Quine, W. V. (1953), 'On a So-called Paradox,' **Mind**, vol. 62.
Quong, J. (2005), 'Disagreement, Asymmetry, and Liberal Legitimacy,' **Politics, Philosophy & Economics**, vol. 4.
Ray, D. and Vohra, R. (2015), 'The Far-sighted Stable Set,' **Econometrica**, vol. 83.
Rosenthal, R. (1981), 'Games of Perfect Information, Predatory Pricing, and the Chain Store,' **Journal of Economic Theory**, vol. 25.
Rubinstein, A. (1989), 'The Electronic Mail Game: Strategic Behavior under Complete Uncertainty,' **American Economic Review**, vol. 79.
Rubinstein, A. (2016), 'A Typology of Players: Between Instinctive and Contemplative,' **Quarterly Journal of Economics**, vol. 131.
Runciman, W. and Sen, A. (1965), 'Games, Justice and the General Will,' **Mind**, vol. 74.
Sarangi, S. (2020), **The Economics of Small Things**, Penguin, New Delhi.
Sartorio, C. (2004), 'How to be Responsible for Something without Causing It' **Philosophical Perspectives**, vol. 18.
Schelling, T. (1960), **Strategy of Conflict**, Harvard University Press, Cambridge, MA.
Scriven, M. (1951), 'Paradoxical announcements,' **Mind**, vol. 60.
Smith, A. (1776), **An Inquiry into the Nature and Causes of the Wealth of Nations**, (Clarendon Press edition, 1978).
Sugden, R. (1986), **The Economics of Rights, Cooperation and Welfare**, Palgrave Macmillan, Houndmills, UK.
Sugden, R. (1995), 'A Theory of Focal Points,' **Economic Journal**, vol. 105.
Schwenkenbecher, A. (2019), 'Collective Moral Obligations: 'We-reasoning' and the Perspective of the Deliberating Agent,' **The Monist**, vol. 102.
Thrasher, J. and Vallier, K. (2015), 'The Fragility of Consensus: Public Reason, Diversity and Stability,' **European Journal of Philosophy**, vol. 23.
Tullock, G. (1971), 'The Paradox of Revolution,' **Public Choice**, vol. 11.
Vanderschraaf, P. (2006), 'War or Peace? A Dynamical Analysis of Anarchy,' **Economics and Philosophy**, vol. 22.
Vanderschraaf, P. (2008), 'Game Theory Meets Threshold Analysis: Reappraising the Paradoxes of Anarchy and Revolution,' **British Journal for the Philosophy of Science**, vol. 71.
Vanderschraaf, P. (2019), **Strategic Justice: Conventions and Problems of Balancing Divergent Interests**, Oxford University Press, New York.
Vanderschraaf, P. (2021), 'Contractarianisms and Markets,' **Journal of Economic Behavior and Organization**, vol. 181.

Von Neumann, and Morgenstern, O. (1944), **Theory of Games and Economic Behavior**, Princeton University Press, Princeton.

Walras, L. (1874), **Elements d'economie politique pure** [4th edition in 1900. English translation by W. Jaffe, in 1954.]