DIALOGUE AND UNIVERSALISM

JOURNAL OF THE INTERNATIONAL SOCIETY FOR UNIVERSAL DIALOGUE

Vol. XXIX **No. 1/2019**

PHILOSOPHY IN AN AGE OF CRISIS: CHALLENGES AND PROSPECTS

PART I
THE SOCIO-POLITICAL SPHERE OF THE HUMAN WORLD

Published three times a year by
INSTITUTE OF PHILOSOPHY AND SOCIOLOGY OF THE POLISH ACADEMY OF SCIENCES and PHILOSOPHY FOR DIALOGUE FOUNDATION

PL ISSN 1234-5792

DIALOGUE
AND
UNIVERSALISM

1/2019

The annual subscription (paper copies) rates are:

Individuals – 50 EUR
Institutions – 70 EUR

Single copies are available at 18 EUR each for individuals and 27 EUR for institutions; some back issue rates are available on request.

All correspondence concerning the subscription of paper copies should be sent to:
Dialogue and Universalism Office: dialogueanduniversalism@ifispan.waw.pl
or to: Ars Polona: agnieszka.morawska@arspolona.com.pl
(regular mail: Ars Polona S.A., ul. Obrońców 25, 03–933 Warszawa, Poland)

Dialogue and Universalism electronic PDF copies are distributed by:
Dialogue and Universalism Office, email: dialogueanduniversalism@ifispan.waw.pl;
the Philosophy Documentation Center: **order [@] pdcnet.org;** and by Central and Eastern European Online Library (**CEEOL)**

All editorial correspondence and submissions should be addressed to:
e-mail: dialogueanduniversalism@ifispan.waw.pl
(regular mail: Dialogue and Universalism, Institute of Philosophy and Sociology of the Polish Academy of Sciences, Nowy Świat 72, 00–330 Warszawa, Poland)

More information on *Dialogue and Universalism* may be found on its website:
http://www.dialogueanduniversalism.eu

Printed by Drukarnia Paper & Tinta, Warszawa

DIALOGUE AND UNIVERSALISM
No. 1/2019

PHILOSOPHY IN AN AGE OF CRISIS: CHALLENGES AND PROSPECTS

PART I
THE SOCIO-POLITICAL SPHERE OF THE HUMAN WORLD

DIALOGUE AND UNIVERSALISM
No. 1/2019

Necip Fikri Alican

FOOL ME ONCE, SHAME ON YOU, FOOL ME TWICE, SHAME ON ME: THE ALLEGED PRISONER'S DILEMMA IN HOBBES'S SOCIAL CONTRACT

ABSTRACT

Hobbes postulates a social contract to formalize our collective transition from the state of nature to civil society. The prisoner's dilemma challenges both the mechanics and the outcome of that thought experiment. The incentives for reneging are supposedly strong enough to keep rational persons from cooperating. This paper argues that the prisoner's dilemma undermines a position Hobbes does not hold. The context and parameters of the social contract steer it safely between the horns of the dilemma. Specifically, in a setting as hostile as the state of nature, Hobbes's emphasis on self-interest places a premium on survival, and thereby on adaptability, which then promotes progressive concessions toward peaceful coexistence. This transforms the relevant model of rationality from utility maximization to utility satisficing, thus favoring the pursuit of a mutually satisfactory outcome over that of the best personal outcome. The difference not only obviates the prisoner's dilemma but also better approximates the state of nature while leaving a viable way out.

Keywords: Hobbes; social contract theory; state of nature; civil society; prisoner's dilemma.

1. INTRODUCTION

Can Hobbesian contractors facing the prisoner's dilemma establish civil society without a regress to the state of nature?[1] This question has fascinated Hobbes scholars at least since David P. Gauthier popularized its formulation in

[1] References to Hobbes are to the World's Classics edition of *Leviathan* (Oxford: Oxford University Press, 1996), giving the page numbers of the "Head" edition (London: Andrew Crooke, 1651) followed by the corresponding part, chapter, and paragraph numbers, which can be tracked through any edition.

terms of game theory.[2] It has also inspired the affirmative answer developed in this paper.

Hobbesian agents contract out of the state of nature and into civil society for the sake of self-preservation and out of the fear of death. But would they continue to abide by that agreement if they thought they could get away with violating it for personal gain? That is the question. The problem is to determine whether such contractors have sufficient reason and motivation to keep the covenants they had sufficient reason and motivation to make. This paper argues that the prisoner's dilemma blocks neither the creation nor the sustenance of the commonwealth in Hobbes.

Setting up the argument requires some reconstruction, though the main constructs are already familiar from the standard exposition of the problem in the literature.[3] The next section employs analytic and tabular paradigms of the challenge to cooperation-in-conflict to illustrate how the prisoner's dilemma is supposed to affect social contract theory in general and why it allegedly concerns the Hobbesian contract in particular. The third section takes up the plausibility, apart from the practice, of applying the prisoner's dilemma to social contract theory. The finding there is that social contract scenarios are not necessarily or automatically susceptible to the prisoner's dilemma, whose relevance varies with the contextual parameters of the case under consideration, including the definition of rationality and the rules of the game. The fourth section demonstrates that the prisoner's dilemma, inapplicable in its classic formulation without modification, is innocuous upon proper reformulation. The basic premise there is that the game and players of the traditional dilemma are fundamentally different from the context and agents of the Hobbesian state of nature as well as from those of the Hobbesian commonwealth.

The differences identified in the fourth section strip the prisoner's dilemma of its disruptive powers over the social structure it is purported to preclude or destroy. From that point on, drawing the conclusion that the dilemma is

---

[2] The application of game theory to Hobbes's social and political philosophy starts with David P. Gauthier, whose analysis in *The Logic of Leviathan* (1969), followed by a series of complementary contributions (1979; 1986; 1988; 1990), inaugurated this mode of treatment and inspired comparable work by others.

[3] Leading examples of the prisoner's dilemma as it is employed in Hobbesian social contract theory include, in addition to the Gauthier entries in the preceding note (n. 2 above), the work of Jean Hampton (1985; 1986), Gregory S. Kavka (1983; 1986), and Edna Ullmann-Margalit (1977). Michael Taylor (1976/1987) merits special mention as he is concerned neither only nor primarily with Hobbes, despite making a significant contribution to the question as it pertains to Hobbes. The list can be extended indefinitely, but it already captures what is at the pioneering forefront of scholarship. That said, a provocative alternative to the standard interpretation of Hobbes with respect to the prisoner's dilemma can be found in Sharon Anne Lloyd (1992; 2009; 2010), who ;discerns a greater motivational force, and therefore superior explanatory power, in what she calls "transcendent interests" (religious and moral interests that transcend the fear of death) than in the psychological factors traditionally adduced to account for and resolve conflict under social conditions representative of Hobbes.

not a compelling problem for Hobbes is a matter of connecting the dots in the reconstruction in progress. The conclusion itself is not a novel one, but the argument is. After all, not everyone connects dots the same way, nor even works with the same ones.[4]

2. PROBLEM

The prisoner's dilemma is a thought experiment in game theory.[5] A popular application is the explication of problems in the social contract tradition. The

---

[4] Mainstream contributions readily demonstrate how different the game can be for each referee: Gauthier, both in *The Logic of Leviathan* (1969) and elsewhere, employs a one-shot (single-play) prisoner's dilemma as opposed to an iterated (repeated) variant, championing what he calls "constrained maximization," essentially utility maximization with normative constraints. That makes his position one of the most ambitious among those who use the prisoner's dilemma to explicate the Hobbesian evolution of civil order. This is because he attempts to get Hobbes from the state of nature to civil society without appealing to the benefits of the kind of experience that comes from repeated interaction with others. His main strategy is a shift from the traditional emphasis on rational choice to an emphasis on a rational disposition to choose, which he claims can be tamer than usually imagined, even without introducing a temporal dimension to the game (Gauthier 1969, 82–87; 1986, 169–170; 1990, 266). Hampton (1986) and Kavka (1986) both embrace the temporal dimension cast aside by Gauthier. Hampton (1986, 74–79, 80–89, 182–188) combines the algorithm of an iterated prisoner's dilemma with the assumption of "short-sightedness" to explain why Hobbesian players do not actually see the benefits of cooperation in the state of nature. Kavka (1986, 109–113, 129–136) also adopts the iterated prisoner's dilemma as a backdrop for his solution, listing seven features increasing the rationality of early cooperation (pp. 134–136). While Taylor (1976/1987) is not concerned exclusively with Hobbes, he can be said to precede Hampton (1986) and Kavka (1986) in the formulation of a solution through an iterated dilemma, as he does in fact analyze the Hobbesian problem in considerable detail (Taylor 1987, 126–150), using the notion of a "supergame" (basically a series of one-shot games) as his version of the iterated dilemma. Ullmann-Margalit (1977) relies on social norms to ward off the prisoner's dilemma. She argues that interactive experience, as soon as it reveals the possibility and benefits of cooperation, begins to reinforce cooperative behavior, which, through natural selection, eventually emerges as a social norm (or rather as a multitude of social norms), which, in turn, continues to defeat or evade the prisoner's dilemma. She does, however, add the threat of punishment to the selection process, thus working with norms backed by sanctions (Ullmann-Margalit 1977, 22, 28).

[5] The prisoner's dilemma owes its name and structure to Albert W. Tucker's presentation of a thought experiment to an audience of psychologists at Stanford University in May of 1950. But the underlying game-theoretic framework had already been developed by mathematicians Merrill Flood and Melvin Dresher of the RAND Corporation as part of their research on conflict and cooperation. The background can be found in abundant detail in the historical and expository account of William Poundstone (1992, 8–9, 116–121). The thought experiment itself is about two suspected felons questioned separately under custody, with no communication between the two, and with each encouraged to confess in exchange for a reduced sentence. They must base their decision on the following information without interaction, negotiation, or consultation: (1) If they both confess, they each receive a reduced sentence. (2) If neither one confesses, they each receive a minor sentence more advantageous than the deal for a double confession. (3) If only one of them confesses, the other one faces the maximum penalty while the confessor goes free. This is the basic scenario, though presentations are usually quantified with a payoff matrix specifying

simplest instantiation is the examination of outcomes of covenanting in a setting with two players each of whom has a choice between two courses of action. Having entered into a covenant, each player must decide whether to keep it or to break it: "adherence" vs. "violation" (or "cooperation" vs. "defection" in more general terms abstracted from the context of the social contract). The main constraint is the absence of communication between the players, whereby the decision of each affects the other, while negotiation and coordination are not possible. The aim of the application is to determine the move a rational player must and therefore would make under the circumstances. The basic setting supports four possible outcomes:

— mutual adherence: both A and B keep the covenant;
— mutual violation: neither A nor B keeps the covenant;
— unilateral adherence: A keeps the covenant while B breaks it;
— unilateral violation: A breaks the covenant while B keeps it.

Critics using the prisoner's dilemma to evaluate the Hobbesian transition from the state of nature to civil society reason as follows:[6]

(P1) Reason dictates that players make a covenant if and only if they prefer mutual adherence (civil society) to mutual violation (the state of nature).
(P2) Reason dictates that players keep a covenant if and only if they prefer mutual adherence (civil society) not just to mutual violation (the state of nature) but also to unilateral violation (personal advantage) and to unilateral adherence (personal disadvantage).
(P3) Reason dictates that players in fact prefer unilateral violation to mutual adherence to mutual violation to unilateral adherence.
(C) Reason dictates that players break the covenant because conditions for entering into a covenant obtain but conditions for keeping a covenant do not.

precise prison sentences to facilitate comprehension and discussion. Variations on the theme are common and options abound for further study: R. Duncan Luce and Howard Raiffa (1957, 94–102) stick with the original formulation in their introduction to game theory in one of the leading textbooks from the heyday of mathematical research into human rationality. Richard Mark Sainsbury (1995, 66–72) can be consulted profitably for a lucid and stimulating discussion of the standard account of the prisoner's dilemma in his influential study of paradoxes. Robert Axelrod's (1981; 1984; 1987) work on cooperation explores the possibilities in employing computer simulations to evaluate strategies for success within a prisoner's dilemma.

[6] The prisoner's dilemma is not the only game in town. At least some commentators using game theory to analyze the Hobbesian context tend to express misgivings and recommend alternatives. Dissenters include Andrew Alexandra (1992), Noel Boulting (2005), Daniel Eggers (2011), Michael Moehler (2009), and Pärtel Piirimäe (2006).

The idea here is that actual circumstances accommodate covenant-making conditions but not covenant-keeping conditions. Rational players prefer mutual adherence to mutual violation but not to unilateral violation. The following chart illustrates individual preferences regarding possible outcomes:

Outcomes of Covenanting		Ranking of Preferences	
Mutual Adherence	A adheres B adheres	A #2	B #2
Mutual Violation	A violates B violates	A #3	B #3
Unilateral Adherence	A adheres B violates	A #4	B #1
Unilateral Violation	A violates B adheres	A #1	B #4

Consider the psychology of rationality implicit in the traditional prisoner's dilemma: Mutual adherence and mutual violation rank, respectively, second and third for each player, whereas unilateral violation and unilateral adherence rank, respectively, first and fourth for each. Since unilateral violation by one player is unilateral adherence by the other, the best possible outcome for one player is the worst possible outcome for the other. Given that the players do not know and cannot affect each other's decisions, violating the covenant is the most advantageous strategy for each, because it helps attain unilateral violation (the most desirable outcome) in case the other player adheres to the covenant, and because it helps avoid unilateral adherence (the least desirable outcome) in case the other player violates the covenant.

Exponents of the prisoner's dilemma in social contract theory contend that Hobbesian agents cannot reasonably adhere to a covenant to create a commonwealth. They hold, as established in the conclusion of the argument diagrammed above, that rationality requires players to break the covenant in all cases.[7]

[7] This conclusion comes with an implicit assumption that people always act in accordance with the requirements of rationality. While the assumption is debatable, that debate is best taken up separately, partly because it is already in a field of its own, but mostly because winning the debate would not free Hobbes of the prisoner's dilemma. That would require establishing not just that people do not always do what they have reason to do, but either more strongly that they never do what they have reason to do (which is patently false) or more specifically that they typically do not do what the prisoner's dilemma says they would do (which is then no longer about rationality in general).

3. ANALYSIS

A natural albeit naïve response to the prisoner's dilemma in the context of the social contract tradition is that a covenant good enough to make should be good enough to keep. This instinctive reaction presupposes that any rationale for entering into a covenant must also work for adhering to that covenant. The simplicity of the assumption, apparently an error, exposes a series of fundamental questions: How does a desirable covenant become an undesirable one? Are people not compelled to keep covenants for the same reasons they are motivated to make covenants? Do people who break a covenant not end up facing the same undesirable circumstances that compelled them to make that covenant?

The last question best emphasizes the practical concern behind the naïve response. If people break a covenant, they are indeed back to square one, including all the unpleasant business that compelled them to make that covenant, but this follows only if all the parties renege. The naïve response assumes that, if anyone breaks the covenant, everyone breaks the covenant. It ignores the gray area between the complete absence and sweeping presence of infractions. This gray area of possible outcomes allows for a prudential calculus of rational expectations and risk tolerance. The prisoner's dilemma thrives on precisely what the naïve response ignores: the potential advantage of unilateral violation in the absence of retaliation. Contractors are forever tempted to violate the covenant and thereby to risk retaliation. Insofar as the violation is unilateral, the violator (defector) reaps the benefits.

The challenge to social contract theory is the overarching requirement for contracting agents to have sufficient reason and motivation to keep the covenants they had sufficient reason and motivation to make.[8] Regarding the inception of the process, here is how Hobbes contemplates the convergence of reasons and motives toward contractual cooperation:[9]

---

[8] I tend to use "motive" and "motivation" interchangeably, since there is a sense in which they mean the same thing: a reason for doing something. I also tend to use them differently, since there is a sense in which motivation is more than just a reason for doing something, indicating instead, or in addition, a cognitive response to that reason, assimilated as a state of mind with the participation of the passions under the direction of the will. As for motives versus reasons, where a motive is a kind of reason to begin with, the difference is that reasons are external whereas motives are internal. Put simply, motives are internalized reasons. When I come in out of the rain, the rain is the reason, my comfort is the motive. A motivation, in contrast to both (in one of its senses), is a mental state, though the word "motivation" can also be used in the sense of "motive." Since I do not assign a special or technical sense to any of these words ("reason"; "motive"; "motivation"), I trust that my usage presents no obstacles to clarity.

[9] The etymology of the word "motivation" precludes its occurrence in Hobbes. Dictionaries date the word back to the late nineteenth century, easily two hundred years after the prime of Hobbes. With no recourse to motivations, Hobbes is instead restricted to motives, which come up in four places in *Leviathan* (1651): 49=1:11:14; 66=1:14:18; 172=2:29:16; 182=2:30:22. His distinction between "the motive" and "end" (66=1:14:18) and his separation of "causes" and

> "The passions that incline men to peace, are fear of death; desire of such things as are necessary to commodious living; and a hope by their industry to obtain them. And reason suggesteth convenient articles of peace, upon which men may be drawn to agreement. These articles, are they, which otherwise are called the Laws of Nature: whereof I shall speak more particularly, in the two following chapters" (Hobbes, 1651, 63=1:13:14).

These are the considerations relevant to the creation of the commonwealth in Hobbes. Peace is a sufficient reason, and self-preservation a sufficient motive, for making covenants. As for keeping covenants, the concerns in making them must be supplemented by the threat of punishment as a necessary condition, whereby nothing else can be sufficient either as a reason or as a motive:

> "For the laws of nature (as *justice*, *equity*, *modesty*, *mercy*, and (in sum) *doing to others, as we would be done to*,) of themselves, without the terror of some power, to cause them to be observed, are contrary to our natural passions, that carry us to partiality, pride, revenge, and the like. And covenants, without the sword, are but words, and of no strength to secure a man at all" (Hobbes 1651, 85=2:17:2).

The discontinuity in conditions for making and keeping covenants suggests that reasons and motives sufficient for making covenants are not sufficient for keeping them. Otherwise, sovereign power would be redundant in maintaining and preserving the commonwealth.

The mere existence of this discrepancy is not a decisive threat to the commonwealth. Hobbes is demonstrably responsive to that threat, the neutralization of which requires only that covenants be kept, not that they be kept for the same reasons and motives for which they were made. Yet the prisoner's dilemma renders that response inadequate. The allegation is not just that the covenant will not be kept for the same reasons and motives that it was made but that it will not be kept at all, despite the institution of punishment. The sovereign cannot fix the dilemma for us.

It may be objected that the prisoner's dilemma is not about what happens in the commonwealth but about the prospects of getting there at all from the state of nature. This is to object that the prisoner's dilemma keeps cooperative progress from ever reaching the point of establishing an institution of punishment. The gist of the claim opposed, however, is neither that the prisoner's dilemma is

---

"motives" (182=2:30:22) both indicate an internalization of motives, proposed in the preceding note (n. 8 above) as the defining difference between motives and reasons. His association of "the power of conduct and command" with "the motive faculty" (172=2:29:16) demonstrates his acknowledgment of an active faculty dedicated to motives, which confirms that Hobbes clearly recognizes the psychological dimension of motives as against the passive logic of reasons. See n. 17 below for further discussion.

limited to the commonwealth nor that sovereign power is sufficient to hold the commonwealth together but merely that Hobbes is aware of the difference between covenant-making conditions and covenant-keeping conditions. The equalizer is laid out in the next section. Yet it is worth noting, even at this point, that the objection reveals a narrow view of punishment as a legal privilege reserved for sovereign power. Punishment without authority, even if it should no longer be called punishment in the absence of legitimacy, is quite apposite to the state of nature where participants quickly learn that people do not respond well to mistreatment, or to outright hostility, or to the constant threat of either. Punishment need not be lawful to be effective, just as it need not be unlawful to be ineffective.

Nevertheless, the sword of the sovereign promises little more than a revision of the risk structure inherited from the state of nature. Granted, once contracting agents exit the state of nature, they enter a state with a different payoff matrix than the one that used to offer net benefits to aggression. But the introduction of lawful punishment does not solve the dilemma, which can be reformulated to accommodate the change in the payoff matrix in civil society: Self-interested contractors who would have violated the covenant in the absence of the sovereign's sword will still violate the covenant if they believe they can do so with impunity. While the sovereign does discourage civil disorder, the sword does not preclude violations of the covenant so long as violators are willing to risk punishment or able to resist it.

Be that as it may, the burden on the sovereign is not as great as it is made out to be. Even if the sword fails to deter all violations of the covenant, individual disturbances do not necessarily add up to sociopolitical chaos on the order of a regress into the state of nature. Just as conditions unique to the state of nature encourage the making of a covenant, forces inherent in civil society discourage the breaking of that covenant.

Accounting for the prisoner's dilemma in civil society is like accounting for moral backsliding in ethical theory or for free-rider problems in economic theory. Moral backsliding is conceivable in any normative ethical theory in the sense that such theories do not come with prevention mechanisms. They are all susceptible to moral backsliding. Yet the threat of moral backsliding does not make morality impossible. It precludes neither moral discourse nor the possibility of a moral life. Likewise, pure externalities do not uproot the economic system: Pure externalities can support only so many free-riders. Beyond a certain volume of utilization, they cease to offer net benefits to free-riders. Instead of collapsing under free-rider problems, the economy works itself back into equilibrium. The analogy with the prisoner's dilemma is that sociopolitical disorder is not a straightforward function of individual violations of the covenant.

This is not to say that individual violations of the covenant cannot possibly add up to a regress to the state of nature. They can, at least in theory, but random violations will not necessarily combine to produce the same effect as mass

movements built on collective or interdependent action. The prisoner's dilemma can accommodate no more than diminishing marginal returns to repeated violations of the covenant. Individual transgressions hold increasingly limited advantages as the prospects for retaliation undermine the possibility of unilateral violation. Defectors cannot enjoy net benefits repeatedly, certainly not indefinitely. And the repetition in question need not come from the same party. Any confrontation between any two parties is a learning experience for everyone in the community.

How many contractors must break the covenant to bring about a regress from civil society to the state of nature? Given the institution of punishment, the question is about breaking the covenant with impunity. This suggests either that the violation is not detected and the violators are not caught or that the violation is detected but the violators are powerful enough to resist punishment. In the first case, a regress is out of the question, because the violation is not serious enough to be detected. In the second case, a regress must have already taken place, because indomitable resistance and overwhelming insurgency are characteristics of the state of nature and not of the Hobbesian commonwealth. The first case represents moral backsliding and free-rider problems. The second case represents anarchy. Neither is under sovereign control. But there are other forces at play.

The social contract is an instrument of change. People come together to change the present state of affairs. Covenants are made in the state of nature but kept or broken in civil society. The decision to accept or reject them is governed, therefore, by conditions in the state of nature, whereas the decision to keep or break them is governed by conditions in civil society. During the transition, an evolution of reasons and motives shifts the sociopolitical focus from individual and independent action in the state of nature to collective and interdependent action in civil society. In the state of nature, the standard of reason by which actions are evaluated, as well as the motivation behind those actions, is individual self-interest manifested as self-preservation. In civil society, self-preservation as the standard of reason in independent action leaves its place to peace as the standard of reason in interdependent action. Meanwhile, individual self-preservation as the ultimate motivation behind independent action leaves its place to collective self-preservation as the ultimate motivation behind interdependent action. Preserving civil society and maintaining social order thus become essential to self-preservation, wherefore potential deterioration through the collapse of civil society emerges as a sufficient reason for adherence and compliance, while the sovereign as enforcer contributes the fear of punishment as an additional motivation for adherence to covenants and compliance with laws.

On the other hand, sociopolitical transformation is not so much a demonstration of the stability of the covenant as it is a condition of that stability. Any appeal to it without argument may well beg the question by simply assuming regress from civil society to the state of nature to be prevented by the transfor-

mation of reasons and motives in the process of transition from the state of nature to civil society. The original charge, to be sure, is not that such a regress can take place despite the structural stability of a political state that has completed the requisite transformation of reasons and motives. It is rather that the transformation invoked either cannot be completed or will not be effective. Thus, the real question concerns the possibility and effectiveness of the very transformation that is supposed to preclude the prisoner's dilemma. The appeal to sociopolitical transformation is still a good response, but it requires showing that and explaining how the transformation can take root firmly enough to define a new and sustainable status quo replacing the state of nature.

The problem is that covenants are only as stable as the interests they serve. If contracting agents enter into covenants only to maximize their self-interest, then they can be expected to keep covenants only insofar as they continue to believe it remains in their own best interest to keep them. They can therefore be expected in general to break covenants if and when they believe it serves their best interest to break them. The basic threat to cooperation, then, is that contracting agents might have sufficient reason and motivation to break covenants if they believe breaking them promotes their interests and either they believe they can break covenants with impunity (whether through clever avoidance of punishment or through forceful resistance to punishment) or they are willing to risk getting caught and being punished.

With self-interest as the common denominator, defending the social contract against the prisoner's dilemma requires proof of the sociopolitical transformation of reasons and motives between the state of nature and civil society. Even if peace is recognized as a means to self-preservation, and even if peace, in time, becomes an end in itself, individual self-interest is not dissociated from the rationale and motivation of contractors. It remains to be shown either that the conventional drive for peace effectively curbs the natural devotion to self-interest or that the former is fully absorbed as an integral part of the latter.

4. SOLUTION

The notion of Hobbesian contractors struggling with the prisoner's dilemma raises at least two important questions: (1) What game are Hobbesian agents playing and what are the rules of that game? (2) How rational are Hobbesian agents and what is a rational agent to do? Hobbes's move from the state of nature to civil society can be justified either by proving that the dilemma does not apply to his game or by showing that his players can work around the dilemma. This section is devoted to demonstrating that the facts happen to coincide with a little of each.

If the prisoner's dilemma is to be taken seriously as a framework of evaluation for the Hobbesian social contract, both the rules and the players of the clas-

sic dilemma must be modified to match the Hobbesian sociopolitical setting. While just about any challenge can be avoided or weakened through reformulation, the point here is not that the dilemma can be reformulated to vindicate Hobbes's sociopolitical theory but that it must be reformulated to capture that theory in the first place. Otherwise, it remains irrelevant in the long run.

The traditional prisoner's dilemma can certainly keep Hobbesian agents from cooperating, but only at first, as any such dilemma persisting throughout the state of nature is by definition an iterated one providing a learning experience. The only outcomes sustainable for any length of time in the Hobbesian state of nature are cooperation and confrontation, which translate roughly into mutual adherence and mutual violation. Given a chance to retaliate, no Hobbesian player would settle for unilateral adherence, and no Hobbesian player could attain and maintain unilateral violation. The temporal dimension of the state of nature allows for no more than an iterated dilemma where any violation leads to mutual violation. This is because unilateral violation triggers immediate retaliation and thus brings mutual violation. The naïve response considered in the beginning of the previous section does not turn out to be so naïve once the circumstances are clearly identified.

In contrast to the players of the traditional prisoner's dilemma, Hobbesian agents communicate and interact with one another, share a common history, and benefit from past experience in decision-making. Any dilemma is iterated throughout the duration of the state of nature, which is plagued by "continual fear" and "danger of violent death" in a war of all against all, where life is "solitary, poor, nasty, brutish, and short" (Hobbes, 1651, 62=1:13:9). Extended exposure to such adverse conditions is bound to make participants sensitive to the cost of failure in competition, thus placing a premium on survival, and consequently on cooperation, while imposing a natural penalty on confrontation.

No conception of rationality can forever support a war of all against all. Hobbesian agents must be in a significantly different frame of mind toward the end of the state of nature than at the beginning when they first encounter one another as rivals. Upon discovering all participants, including themselves, to be of roughly equal strength and wit (Hobbes, 1651, 60–61=1:13:1–2), they must at some point learn to temper the complex psychology of competition, diffidence, and glory in appreciation of the peace of mind that comes with self-preservation through peaceful coexistence.

One may be tempted to object that, if the prisoner's dilemma is fully iterated prior to the decision to covenant, then Hobbesian agents must have already learned through experience to live in some sort of harmony, and therefore that both the covenant and the sovereign are superfluous. But this objection does not help the critic drawing on game theory: If the dilemma is thus iterated to exhaustion, then the game is over, and the contract prevails. Indeed, that makes the iteration itself self-destructive as well, but only in the manner of an analgesic pill that disintegrates in the process of alleviating

pain, hence as the result of a positive outcome. Nor would the critic fare any better objecting that the dilemma is not iterated at all, which is empirically wrong, as it contradicts the natural context where aggression has unmistakable consequences.

Given the failure of the objection that the dilemma might eventually be iterated out of existence, plus the inaccuracy of the opposite objection that it is not iterated at all, the only critical option left is to deny that the dilemma is (or can ever be) sufficiently iterated to inspire cooperation. But that will not work either. The problem with this intermediate alternative is that it holds us captive in a transitional stage between first contact and social contract. Was there ever such a stage? Of course, there was. If there was ever a state of nature, and if there is now civil order, then there was once something in between. But to object that the dilemma can never be sufficiently iterated to get to the social contract stage is to deny precisely what has in fact happened. And it will not do to admit only that the dilemma could eventually be iterated to a sufficient degree, while denying that it has already been iterated to that level, which is as good as denying that we all live in civil society. To affirm the instability and to predict the collapse of civil society is one thing, to deny its existence altogether is quite another. The thesis of this paper is that the dilemma is sufficiently iterated to make it both reasonable and desirable to covenant without regress. The remainder of this section fleshes out the evidentiary basis and supporting rationale together with the details of the iteration process.

The prisoner's dilemma is based on economic theory in its origins.[10] But the attainment of the best possible outcome is no longer an unqualified goal in economics. Utility maximization is an extravagant definition of economic rationality. A more sober paradigm awaits, among other places, in Herbert Alexander Simon's initiative to disown the utility-maximizing agent as the epitome of rationality in economic theory.[11] Simon proposes replacing utility

---

[10] John von Neumann and Oskar Morgenstern (1944) are responsible for popularizing the application of game theory in the field of economics. Richard C. Jeffrey (1965) is famous for offering a systematic approach to probability and rationality in a work that has become a classic in the philosophical foundations of decision theory. Derek Parfit (1984, 3–114), in a study best known for advancing our understanding of personal identity, invokes the prisoner's dilemma and taps into the tools of moral mathematics to expose the weaknesses of ethical theories grounded in self-interest, most notably, ethical egoism. Paul K. Moser (1990) offers a collection of influential essays on rational choice and game theory, with authoritative pieces on the prisoner's dilemma (pp. 271–334).

[11] Herbert Alexander Simon (1916–2001) was a professor of psychology and computer science at Carnegie-Mellon University. His chief academic interests were human rationality, rational choice theory, and artificial intelligence. Drawing on his core expertise to challenge utility maximization as one of the fundamental assumptions of economic theory, in 1978, he received the Alfred Nobel Memorial Prize in Economic Sciences. The basic insight earning him global recognition was that rational people opt for a satisfactory level of utility instead of pursuing the maximum possible utility. Satisficing as a solution originates specifically with two of his papers on the topic: "A Behavioral Model of Rational Choice" (1955) and "Rational Choice and the Structure

maximization as an ideal goal with utility satisficing as a realistic goal: Rational agents are prepared to accept an outcome with a satisfactory level of utility as opposed to holding out for the outcome with the greatest possible utility.[12] Economic rationality rests with the agent who is not merely a utility calculator but a reasonable person in a broader sense.

Three principles stand out in the theory of economic rationality as Simon has it: (1) Rational agents are sensible enough to realize that it is often impractical and unreasonable to sort through all possible outcomes to determine which one provides the greatest utility. (2) Even if the number of possibilities is small enough to allow for the evaluation of all options, rational agents know that the outcome with the greatest utility is not necessarily an outcome they can realistically expect to attain. (3) Even if the outcome with the greatest utility can be identified, and turns out to be realistically achievable, it must be morally acceptable, because rational agents are not cold and prudential utility calculators with no concern for others.

What does all this mean for the social contract envisaged by Hobbes? The first principle does not help avoid a prisoner's dilemma of any kind in the Hobbesian setting, given that players can reasonably be expected to sort through all four outcomes.[13] And the third principle seems hardly convincing as a description of self-preserving and death-fearing Hobbesian agents in the context of competition, diffidence, and glory.[14] However, the second principle points to a way out of any such dilemma in the state of nature: Rational players, and especially those participating in an iterated dilemma, are reasonable enough to

---

of the Environment" (1956). A broader first-hand account of his views on human rationality is also available in his *Reason in Human Affairs* (1983).

[12] Mine is not the first attempt to invoke satisficing in the context of the prisoner's dilemma. Another is a contribution by Robert E. Goodin (1988), who urges a distinction between "up-side satisficing" and "down-side satisficing." Up-side satisficers are indifferent between good outcomes above a minimally acceptable floor for the positive, while down-side satisficers are indifferent between bad outcomes below a maximally tolerable ceiling for the negative. Strategic decision-making takes place between the floor and the ceiling. Goodin maintains that these thresholds become particularly relevant where "superabundance" or "scarcity" come into play. The Hobbesian war of all against all would seem to require classification under the latter heading, though note that Goodin's thesis is not specifically about Hobbes, who is never even mentioned in the article.

[13] The relevance of the first principle depends on the complexity of the setting. The one here happens to be a simple setting with two players separately deliberating adherence versus violation. The complexity increases with the number of players, and with the opportunity to play against multiple players at once, especially if they are all acting alone. The setting in Hobbes might well be so complex as to validate the first principle, but this is not necessary for the thesis defended here, which would only benefit from the assumption of greater complexity, as opposed to being hindered by it, given that the first principle constitutes one more reason to favor satisficing behavior over maximizing behavior.

[14] Invoking the third principle, as is the case with the first, discussed in the preceding note (n. 13 above), would support rather than undermine the thesis defended here. Again, doing so is not necessary. That said, appeals to this principle are neither unreasonable nor uncommon. Gauthier (1990, 232), for one, goes so far as to insist on "giving economic man a moral dimension."

realize that the most advantageous outcome is not necessarily the most likely outcome. Such players seek a satisfactory outcome that is attainable without much of a struggle instead of pursuing the best outcome that may or may not be attainable no matter the effort.

Any decision procedure based on cost-benefit analysis naturally includes the feasibility as well as the desirability of attaining each outcome. Rational players in an iterated dilemma will eventually (probably rather quickly) realize that attempts to attain unilateral violation will always be (as they always have been) countered by efforts to avoid unilateral adherence. As a result, rational players will come to prefer sustained mutual adherence over momentary or temporary unilateral violation even if they understand and covet the advantages of the utility-maximizing unilateral violation. The decision-making process will be adjusted accordingly to incorporate probabilities connecting relevant experiences with rational expectations.

The crux of the methodological distinction, however, is not that probability calculations are alien to utility maximization and peculiar to utility satisficing as distinct processes, but that the most salient probabilities in this particular case favor a satisficing approach over a maximizing approach. Otherwise, probabilities are employed just as routinely in maximization strategies as they are in satisficing strategies. Any sensible scheme for promoting utility, whether to the maximum level or to an acceptable level, must take stock of the likelihood of outcomes as well as their nominal utilities.

The point of the distinction is to link the right players with the right game: The classic prisoner's dilemma, a one-shot game with no room for iteration, works with utility maximization in a setting where expectations are not informed by past experience with retaliatory tendencies, as there is no such continuity when the game begins and ends all at once in simultaneous actions implementing final decisions. This is a vacuous configuration completely ignoring the past and either denying or at least distorting the future. At the opposite extreme, where time extends in both directions, the Hobbesian state of nature affords ample room for iteration experience to cultivate utility satisficing as the dominant strategy. This is a more realistic configuration grounding the likelihood of each outcome in the pattern of responses the contemplated actions have been known to elicit from the other players, currently all engaged in vigilant observation.

No model of the state of nature can capture its essence without acknowledging a temporal dimension, one of indefinite duration, not just backward in time where players can draw on past experience, but also forward in time where they are stuck with each other for the foreseeable future. The problem with the traditional one-shot game is not just that it ignores past interaction but also that it rules out a common future. That being so, the conceptual insight of a one-shot game is restricted to a snapshot of the state of nature at a random instant misrepresenting, or at best inadequately reflecting, the natural tendencies of its partici-

pants in the long run. What is required instead is a comprehensive chronicle of the psychological and sociological foundations of the state of nature throughout its duration.

Satisficers in an iterated dilemma are not necessarily smarter or wiser or more logical than maximizers in a classic dilemma. All rational players, no matter what game they are playing, can reasonably be assumed to be capable of factoring in the possibility of retaliation insofar as everyone knows what retaliation is and understands what tends to bring it out in people. The problem is not so much with the players as it is with the game. The iterated version of the prisoner's dilemma illustrates the Hobbesian context better than the classic version, because the experience inherent in iteration makes retaliation a prime concern, and its anticipation a strategic consideration, whereas the absence of a mutual future in the one-shot game makes all such anticipation irrelevant, whether or not the players in the one-shot game would actually be able to anticipate it despite the absence of a mutual past.

The future in typical circumstances of conflict-versus-cooperation in the real world can hardly be expected to unfold as smoothly as depicted in the classic prisoner's dilemma where everyone goes their separate ways after a momentous decision never to be revisited. We all have to live with our decisions, and worse, with the people affected by our decisions. No one would confess under interrogation if the incentive to do so were strictly monetary, with the confessor being awarded, say, a million dollars, but having to share a prison cell with the other suspect-cum-convict. Even if we assume that the million dollars, or whatever amount works best, is a sufficient incentive for either suspect to forgo their freedom, the prospect of their continued interaction will be an even stronger deterrent for each. This is to say nothing of the notorious prison stigma attached to informants, assuming that the confession is public knowledge, which it almost certainly will be.[15] The moral agents Hobbes guides out of the state of nature likewise come from a closely interactive scenario of open and fierce competition as opposed to the safety of a controlled environment where people do not have to deal with each other after getting what they want at the expense of each other. The payoff matrix becomes irrelevant if one has to live with Keyser Söze after ratting him out.

To recapitulate the main lines of the operating difference: participants in the traditional prisoner's dilemma are utility-maximizing players acting upon nominal utilities in a payoff matrix formulated without the benefit of a learning curve shaped by past experience with retaliation, whereas Hobbesian contrac-

[15] One way to maintain the privacy of unilateral violations in the modified model (the financial hybrid) may be to assign the same prison term to a single confession as to a double confession. The consequent disruption of the incentive differentials inherited from the original model can presumably be compensated through further financial arrangements. Yet the fine-tuning of conditions is a convenience more common in thought experiments than in the real world where unilateral violations hardly ever go undetected or unpunished.

tors are utility-satisficing players acting upon expected utilities indexed to different probabilities associated with various expectations grounded in vast experience with retaliation.[16]

The following chart illustrates the effects of relaxing the criterion of rationality from utility maximization to utility satisficing in an iterated dilemma:

Outcome	Nominal Utility	Maximizing Decision	Probability	Expected Utility	Satisficing Decision
Mutual Adherence	50 utils	rank #2	50%	25 utils	rank #1
Mutual Violation	0 utils	rank #3	50%	0 utils	rank #2
Unilateral Adherence	–100 utils	rank #4	0%	0 utils	rank #2
Unilateral Violation	100 utils	rank #1	0%	0 utils	rank #2

Utility-maximizing players and utility-satisficing players have the preference patterns depicted in the third and sixth columns respectively. The utility-maximizing players of the classic prisoner's dilemma prefer the outcome with the greatest nominal utility, which here reflects the absence of a temporal dimension, and thereby a complete lack of experience with iteration, together with a correlative lack of concern with retaliation. The utility-satisficing players in

[16] This is a distinction specifically between the players of the traditional prisoner's dilemma in its simplest setting and the contracting agents invoked in the Hobbesian sociopolitical context with iteration. It is not a general distinction between utility maximization and utility satisficing. To elaborate, I am not saying that probabilities corresponding to rational expectations are, as a rule, always relevant to satisficing and never so to maximization. I am saying only that there are no learned probabilities to associate with any shared experiences in the traditional one-shot, two-player setting where players are typically construed as seeking to maximize utility in an isolated decision. This is because there is no shared experience to draw on, and no mutual future to project, where there is no iteration to carry out. While this note is largely a repetition of caveats in the main text, the disclaimers there failed to keep a critic from detecting, and protesting in private communication, an allegedly sweeping and therefore illicit move to associate probabilities strictly with utility satisficing while denying their relevance to utility maximization. The basic premise, in contrast, is that iteration creates shared experiences which are then projected into the future as probabilities to consider. Unprofitable past interaction, when repeated often enough with the same results, turns Hobbesian players into satisficers. The underlying explanation is not that probabilistic reasoning is a formal requirement of satisficing strategies, while working with probabilities is a methodological abomination in maximizing strategies, but that the inspiration for the satisficing approach in this particular case happens to come with relevant experiences that lend themselves to useful strategic forecasting. None of that informs a thought experiment frozen in the timeframe of a single decision without a past or a future.

Hobbes prefer the outcome with the greatest expected utility, which is thoroughly conditioned by past experience with countermoves and strongly indicative of the possibility of more of the same. Since the outcome with the greatest nominal utility is unilateral violation, that is what the utility-maximizing players prefer under these circumstances. But utility-satisficing players reason that others are ready, willing, and able to retaliate, which then pushes unilateral violation beyond reasonable expectations. They therefore prefer the outcome with the greatest expected utility, which is mutual adherence.

The assignment of values in the chart is flexible but not arbitrary. Concerning the nominal utilities of outcomes, the fact that the players exit the state of nature shows that it has no utility for them. Hence, the state of nature, for which zero utils seems to be a fitting assignment, can be taken as a point of reference for estimating the utility of the outcomes of covenanting. Mutual violation rates zero utils because that outcome, in effect, takes everyone back to the state of nature. Unilateral adherence rates negative one-hundred utils because that outcome is not merely unsatisfactory but insufferable as well. Unilateral violation rates one-hundred utils, and mutual adherence fifty utils, because the former is more desirable than the latter, while both are more desirable than the other two outcomes.

But what if the numbers are wrong? Is the state of nature, for example, represented accurately? Is zero a reasonable approximation for the perceived value of the state of nature? Is that figure, zero, even an approximation, or just an arbitrary designation with no basis in realty? Perhaps the assumption of zero value underestimates the inherent misery in the state of nature, which can only be captured with a negative value. Or alternatively, perhaps it is not all that miserable and holds some value, however small, which would then require a positive assignment. In either case, allocating zero utility to the state of nature would be not just erroneous but also misleading, especially so because the utility of the state of nature is a benchmark for value assignments for the outcomes of covenanting.

Regardless of any apparent cause for suspicion, the proper balance of utilities is not a matter of quantitative precision. The numbers are grounded in qualitative considerations: The state of nature and mutual violation have the same value, whatever that may be, but mutual violation cannot have a positive value great enough to alter its position in a ranking of preferred outcomes. The assignment of positive, neutral, and negative values is likewise consistent with the standard assumptions of game theory: Mutual adherence and unilateral violation have positive values, while mutual violation has a neutral value and unilateral adherence has a negative value. None of these assignments represents a fanciful estimate.

What is definitive in the chart is the basic assumptions, not the representative numbers. As long as the principles are preserved, the numbers can be modified without affecting the outcome. The results do not depend on the utilities and

probabilities chosen. To be blunt, the tabular presentation is not rigged to make things come out right. As a matter of fact, the same chart can be reconstructed in terms that do not involve numbers for utility assignments. We can just as easily work with parameters that have no particular numerical value, so long as the logical and quantitative relationships between the parameters continue to represent reasonable expectations. The result will be the same, with the maximizing decision favoring unilateral violation and the satisficing decision favoring mutual adherence:

Outcome	Nominal Utility	Maximizing Decision	Probability	Expected Utility	Satisficing Decision
Mutual Adherence	x utils	rank #2	50%	x/2 utils	rank #1
Mutual Violation	0 utils	rank #3	50%	0 utils	rank #2
Unilateral Adherence	–z utils	rank #4	0%	0 utils	rank #2
Unilateral Violation	x+y utils	rank #1	0%	0 utils	rank #2

The probability of attaining each outcome is based on the nature of iteration. In an iterated dilemma, a final outcome involving a unilateral advantage or disadvantage has a probability of zero, since the players are able to retaliate. And they are able to retaliate effectively, because Hobbesian agents, as already mentioned, are of roughly equal strength and wit (Hobbes 1651, 60–61=1:13:1–2), meaning that any one of them can either directly or indirectly harm any other. This precludes the realization of the best and worst possible outcomes, unilateral violation and unilateral adherence. The remaining outcomes, mutual adherence and mutual violation, seem equally likely. So each merits a probability assignment of fifty percent. The elimination of unilateral outcomes makes mutual adherence the most plausible outcome, since rational players prefer mutual adherence to mutual violation. Probabilistic reasoning may admittedly be just as relevant in connection with other considerations that may occur to players engaged in cost-benefit analyses, but any such reasoning outside the present concern with iteration experience will have the same effect, if any, on both maximizing and satisficing players.

But does grounding the plausibility of an outcome in its desirability beg the question of its probability? Although the equal probability of mutual adherence and mutual violation is a reasonable assumption, the tenability of the solution developed in this section does not turn on that point. The demonstration can be

continued without specific probabilities. Given that unilateral outcomes are unsustainable, the probabilities of mutual outcomes are reciprocal: The percentage probability of one mutual outcome's occurrence subtracted from one-hundred percent is the percentage probability of the other mutual outcome's occurrence. Since the value of mutual adherence is positive, whereas that of mutual violation is zero, the expected utility of mutual adherence in these circumstances will always turn out to be greater than the expected utility of mutual violation. The following reformulation of the chart illustrates this point:

Outcome	Nominal Utility	Maximizing Decision	Probability	Expected Utility	Satisficing Decision
Mutual Adherence	x utils	rank #2	a%	(x•a)/100 utils	rank #1
Mutual Violation	0 utils	rank #3	(100–a)%	0 utils	rank #2
Unilateral Adherence	–z utils	rank #4	0%	0 utils	rank #2
Unilateral Violation	x+y utils	rank #1	0%	0 utils	rank #2

The maximizing decision still favors unilateral violation while the satisficing decision continues to favor mutual adherence. This is because the difference is not in the numbers chosen but in the rationale they represent. However tempting unilateral violation may be, especially and perhaps even compellingly so in the beginning, subsequent interaction will effectively prevent the prisoner's dilemma from turning into a permanent obstacle to contracting and coexistence in Hobbes.

5. CONCLUSION

The prisoner's dilemma is ultimately ineffective against Hobbesian social contract theory because the relevant version cannot survive repeated interdependent action, especially not for an indefinitely long duration on a global scale, where the underlying conception of human rationality includes experience and learning among its essential characteristics. Learning from experience guarantees that reasons and motives sufficient for making covenants remain sufficient for keeping them when there is extended interaction.[17] In a Hobbesian setting,

[17] Anyone interested in finer distinctions between reasons and motives, though nothing in my discussion turns on it (cf. ff. 8, 9 above), may consult, among other sources, the contrasting views

people soon learn the difference between short-term benefits and long-term interests, as well as the difference between what is advantageous and what is acceptable, which eventually teaches them to subordinate optimal results to satisfactory results.

The reformulation here of the rules of the game in the traditional prisoner's dilemma, together with the attendant redefinition of the rationality of the players, is a faithful representation of how Hobbesian agents can be expected to behave. It is more reasonable to expect them to adopt civilization, and to adapt to it, as always for self-interest, than to expect them to abandon the security of peace for the prospect of a gain which they can be fairly confident would be not only temporary but reversible as well, possibly with overcorrection. These considerations confirm that the Hobbesian social contract can survive the prisoner's dilemma.

REFERENCES

Alexandra, Andrew. 1992. "Should Hobbes's State of Nature Be Represented as a Prisoner's Dilemma?" *The Southern Journal of Philosophy*, 30, 2 (Summer), 1–16.

Axelrod, Robert. 1981. "The Emergence of Cooperation among Egoists." *The American Political Science Review*, 75, 2 (June), 306–318.

---

of Christine M. Korsgaard (1986, 10), who maintains that reasons and motives for moral action are the same in Hobbes, and Anita M. Superson (2009, 155–158), who argues that Hobbes recognizes their equivalence in the state of nature but sees a divergence in civil society. Note that their disagreement over reasons versus motives is not generalized across all actions but restricted specifically to moral actions. What Korsgaard equates is not the reasons and motives for doing something in general, but the reasons that make actions right and the reasons why we do them, the latter being the motives for doing the right thing. This is an ancient perspective dating as far back as Socrates, who famously insisted, to hear Plato tell it, that to know the good is to do the good (*Meno*, 77a–78b; *Protagoras*, 345d–347a; *Republic*, 589c), though Plato seems just as comfortable promoting the precept without Socrates (*Timaeus*, 86d–87b; *Laws*, 731c, 860d–861e). The same principle can be found in the Bible (James, 4:17), which may have been even more relevant for Hobbes, perhaps not personally, but certainly in consideration of the times, and consequently of his circumstances, as evidenced by the fact that the subject of God comes up as early as the very first sentence of *Leviathan*. Superson's disagreement with Korsgaard is that, as the Hobbesian agent moves from the state of nature to civil society, the motive stays the same while the reasons change. The motive is still self-interest, but the rationale of what is in one's interest changes in response to the transition, which then either replaces or transforms the reasons for action, thus driving a wedge between justification and motivation, or in other words, between "justifying reasons" and "motivating reasons." My own position on the rationale and rationality of Hobbesian agents is closer to that of Superson. Yet since the difference of opinion between Korsgaard and Superson is in a specifically moral context, where Hobbes is a vehicle for discussion rather than the focus of attention, I must add that my overarching position on the morality of actions is in greater alignment with John Stuart Mill than it is with Korsgaard or Superson (or with Hobbes as interpreted by either of them): "The morality of the action depends entirely upon the intention—that is, upon what the agent *wills to do*. But the motive, that is, the feeling which makes him will so to do, when it makes no difference in the act, makes none in the morality: though it makes a great difference in our moral estimation of the agent, especially if it indicates a good or a bad habitual *disposition*—a bent of character from which useful, or from which hurtful actions are likely to arise" (Mill, 1969, 220n; bracketed variations from other editions omitted).

_____. 1984. *The Evolution of Cooperation*. New York: Basic Books. Revised edition: 2006.
_____. 1997. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton: Princeton University Press.
Boulting, Noel. 2005. "Ought Hobbes's Natural Condition of Mankind Be Represented as a Prisoner's Dilemma?" *Hobbes Studies*, 18, 1 (January), 27–49.
Eggers, Daniel. 2011. "Hobbes and Game Theory Revisited: Zero-Sum Games in the State of Nature." *The Southern Journal of Philosophy*, 49, 3 (September), 193–226.
Gauthier, David P. 1969. *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*. Oxford: Oxford University Press.
_____. 1979. "Thomas Hobbes: Moral Theorist." *The Journal of Philosophy*, 76, 10 (October), 547–559.
_____. 1986. *Morals by Agreement*. Oxford: Oxford University Press.
_____. 1988. "Hobbes's Social Contract." *Noûs*, 22, 1 (March), 71–82.
_____. 1990. *Moral Dealing: Contract, Ethics, and Reason*. Ithaca: Cornell University Press.
Goodin, Robert E. 1988. "Some New Sources of Social Conflict: Transformations of Mixed-Motive Games." *The British Journal of Sociology*, 39, 3 (September), 441–451.
Hampton, Jean. 1985. "Hobbes's State of War." *Topoi*, 4, 1 (March), 47–60.
_____. 1986. *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press.
Hobbes, Thomas. 1651. *Leviathan*. Edited with an introduction and notes by John Charles Addison Gaskin. World's Classics. Oxford: Oxford University Press, 1996.
Jeffrey, Richard C. 1965. *The Logic of Decision*. New York: McGraw-Hill. Second edition: 1983.
Kavka, Gregory S. 1983. "Hobbes's War of All Against All." *Ethics*, 93, 2 (January), 291–310.
_____. 1986. *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.
Korsgaard, Christine M. 1986. "Skepticism about Practical Reason." *The Journal of Philosophy*, 83, 1 (January), 5–25.
Lloyd, Sharon Anne. 1992. *Ideals as Interests in Hobbes's Leviathan: The Power of Mind over Matter*. Cambridge: Cambridge University Press.
_____. 2009. *Morality in the Philosophy of Thomas Hobbes: Cases in the Law of Nature*. Cambridge: Cambridge University Press.
_____. 2010. "The Moral Philosophy of Thomas Hobbes: A Reply to Critics." *Hobbes Studies*, 23, 2 (December), 180–187.
Luce, R. Duncan, and Howard Raiffa. 1957. *Games and Decisions*. New York: John Wiley and Sons.
Mill, John Stuart. 1969 (1861). *Utilitarianism*. References in this paper give the pagination of the fourth edition (London: Longmans, Green, Reader, and Dyer, 1871) as it appears in *Essays on Ethics, Religion and Society* (203–259), vol. 10 of the *Collected Works of John Stuart Mill*. Robson, John M. (Ed.). Priestly, F. E. L. (Series Ed.). Toronto: University of Toronto Press, 1969.
Moehler, Michael. 2009. "Why Hobbes' State of Nature Is Best Modeled by an Assurance Game." *Utilitas*, 21, 3 (September), 297–326.
Moser, Paul K. (Ed.). 1990. *Rationality in Action: Contemporary Approaches*. Cambridge: Cambridge University Press.
Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
Piirimäe, Pärtel. 2006. "The Explanation of Conflict in Hobbes's *Leviathan*." *Trames: A Journal of the Humanities and Social Sciences*, 10, 1 (60/55), 3–21.
Plato. *Platonis Opera*. Five volumes. Critical texts with introductions and notes. Burnet, John (Ed.). Oxford Classical Texts. Oxford: Oxford University Press, 1900–1907. New edition underway since 1995.
_____. *Complete Works*. English translations by various hands. Cooper, John M. (Ed.). Indianapolis: Hackett Publishing Company.
Poundstone, William. 1992. *Prisoner's Dilemma*. New York: Doubleday.
Sainsbury, Richard Mark. 1995 (1987). *Paradoxes*. Cambridge: Cambridge University Press.
Simon, Herbert Alexander. 1955. "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics*, 69, 1 (February), 99–118.
_____. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review*, 63, 2 (March), 129–138.
_____. 1983. *Reason in Human Affairs*. Stanford: Stanford University Press.
Superson, Anita M. 2009. *The Moral Skeptic*. Oxford: Oxford University Press.
Taylor, Michael. 1976. *Anarchy and Cooperation*. London: John Wiley and Sons. Revised and expanded edition subsequently published as *The Possibility of Cooperation*. Cambridge: Cambridge University Press, 1987.
Ullmann-Margalit, Edna. 1977. *The Emergence of Norms*. Oxford: Oxford University Press.
Von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

ABOUT THE AUTHOR — independent scholar in ethics and metaphysics, both especially from a historical perspective. PhD, Washington University in St. Louis. Author of *Mill's Principle of Utility: A Defense of John Stuart Mill's Notorious Proof* (1994) and *Rethinking Plato: A Cartesian Quest for the Real Plato* (2012) as well as various articles in peer-reviewed academic journals.

E-mail: necipalican@gmail.com

NOTE TO OUR CONTRIBUTORS

Manuscripts (in English) may be considered for *Dialogue and Universalism* if they have not previously been published, except special cases negotiated individually: We may admit papers which have not yet been published in English and when the copyright are given us by their first publishers. We are not inclined to accept double submissions. Authors are not being charged any fees.

Manuscripts of submitted works, in the Word format, should be sent in electronic form to the address: dialogueanduniversalism@ifispan.waw.pl or to: mczarnoc@ifispan.waw.pl. They should be double-spaced, include an abstract of approximately 150 words, key words, and intertextual headings. Footnotes (not endnotes) and/or references should be in a separate section. The first reference to a book or journal article should have complete bibliographical information.

The submitted manuscript should contain information about the author. This should be no longer than up to approximately 100 words. It should include academic degree, scholarly affiliation, membership in important organizations, especially international ones, up to five titles of the author's most significant books or papers with bibliographical data, the author's address, and e-mail.

We also publish books reviews, discussion notes, and essay reviews.

Submissions are sent to two referees for blind peer review. To facilitate blind refereeing, the author's name and address should appear on a detachable title page, but nowhere else in the article. The list of the current referees is available on our website.

The suggested length of *Dialogue and Universalism* articles is to 9 000 words, of reviews and discussion notes—to 2000 words.

Submitted manuscripts should follow the bibliographical specifications presented on our website: dialogueanduniversalism.eu in the form of typical instances.

We encourage gender-neutral and international language wherever possible. We allow both British and American usage, but there must be consistency within the individual article.

Authors of articles published in D&U assign copyright to the journal. If they wish to reuse their publications formal D&U permissions are downloaded free of charge.

The paper version of D&U is the basic and only one.

Authors receive free one copy of the issue containing their article. Additional copies may be obtained at half the regular price.

Our website is: dialogueanduniversalism.eu

Our email addresses: dialogueanduniversalism@ifispan.waw.pl

and: mczarnoc@ifispan.waw.pl

ABOUT *DIALOGUE AND UNIVERSALISM*

Dialogue and Universalism is published since 1973, first, under the title *Dialectics and Humanism,* since 1990 under the title *Dialogue and Humanism. The Universalist Quarterly*, and since 1995 under the present title.

Dialogue and Universalism tends to show that philosophy is an essential eternal domain of human culture and an inevitable element of the nowadays human world. Critical and creative rational thinking is an opportunity for humankind to resist the lies and illusions of ideological manipulations that serve as instruments of enslavement and oppression. This open and broad vision of philosophy as an expression of human rationality offers a chance to free people's awareness, to open their minds, and to extend their possibilities of thinking and acting. In doing so, philosophical reflection is able to refine and renew old ideals and values as well as to create new ones. It is these two aspects—free consciousness and new ideals—that are necessary to build a more decent human world. The International Society for Universal Dialogue (ISUD) community is convinced that philosophy has an important role to play in the struggle for the future of humanity. Philosophy with its amazingly sophisticated ways of thinking disposes a tremendous power to cope with the world and to change it. Philosophy is free from technical and practical interests, and constituted by the pursuit of removing—from a highly distanced and neutral perspective—falsehood, prejudices, mental, cultural, religious and social slavery. So it gives a hope for human beings' emancipation as well as for an alteration of the world.

Dialogue and Universalism is wholly open for all scholars in the world, not being a publishing forum for the ISUD members only. All contributors are equally kindly welcome.

Dialogue and Universalism publishes monothematic issues. However, each monothematic issue of the journal is completed with a few texts thematically different from the main theme of the issue. This decision allows for a broader thematic diversity. The forthcoming main themes are announced in advance at the *Dialogue and Universalism* website. The announcements should be treated as an open invitation for every scholar to participate in *Dialogue and Universalism* projected enterprises. Besides, proposals of themes and contents of next *Dialogue and Universalism* issues are kindly welcome.