Title: Crucial Evidence: Hobbes on Contractual Obligation

Author's Name: Luciano Venezia (Universidad Nacional de Quilmes / Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina)

Abstract: In this paper, I will introduce the notions of crucial argument and crucial evidence in the philosophy of intellectual history (broadly construed, including the history of political thought). In turn, I will use these concepts and take sides in an important controversy in Hobbes studies, namely whether Hobbes holds a prudential or a deontological theory of contractual obligation. Though there is textual evidence for both readings, I will argue that there is especially relevant evidence—crucial evidence—for interpreting Hobbes's account in a deontological fashion.

Key Words: Underdetermination of interpretation, Crucial evidence, Hobbes, Contractual obligation

Crucial Evidence: Hobbes on Contractual Obligation[1]

1. Introduction

In this paper, I will introduce the notions of crucial argument and crucial evidence in the philosophy of intellectual history (broadly construed, including the history of political thought). In turn, I will use these concepts and take sides in an important controversy in Hobbes studies, namely whether Hobbes holds a prudential or a deontological theory of contractual obligation. Though there is textual evidence for both readings, I will argue that there is especially relevant evidence—crucial evidence—for interpreting Hobbes's account in a deontological fashion.

The paper is organized as follows. In section 2, I will discuss underdetermination of interpretation. In section 3, I will introduce the ideas of crucial argument and crucial evidence. In section 4, I will analyze the difference between deontological contracts and self-interested agreements. In section 5, I will show that there is textual evidence to construe Hobbes's account of contractual obligation in prudential as well as deontological terms. In section 6, I will argue that Hobbes's example of the promise to the thief constitutes a piece of crucial evidence. In section 7, I will claim that Hobbes holds a deontological theory of contractual obligation. In section 8, I will bring the paper to a close introducing some final remarks.

2. Underdetermination of Interpretation

[1] I have delivered previous versions of this paper at the Institute of Philosophy, University of Buenos Aires, in April 2012, the Hobbes Workshop, King's College, London, in May 2012, and the Department of Philosophy I, University of Granada, in May 2012. I thank the audiences of these events. I also acknowledge the comments and suggestions of Adrian Blau, Robin Douglass, William A. Edmundson, Pedro Francés Gomez, A. P. Martinich, Johan Olsthoorn, Jon Parkin, John T. Sanders, Peter Schröder, Quentin Skinner, Laurens van Appeldoorn, Paul Weirich and Romina Zuppone.

Although intellectual history deals with logical and conceptual issues, much of the work done in the field involves empirical research.[2] Intellectual history thus possesses many similarities to other empirical endeavors, notably natural and social science.[3] For that very reason, it also has problems and puzzles analogous or at least related to those of science.

One standard problem in the philosophy of science is the "underdetermination of theory by data." Underdetermination of theory by data means that the same empirical evidence is compatible with two or more mutually incompatible scientific theories. Intellectual history has an analogous problem: "underdetermination of interpretation by textual evidence." Underdetermination of interpretation by textual evidence means that a text or set of texts provides evidence for two or more incompatible readings.

The idea that there are mutually incompatible scientific theories that entail the same empirical evidence is an "obvious truth" of the philosophy of science.[4] Although it does not seem equally obvious that there is such a thing as underdetermination of interpretation by textual evidence,[5] it seems true as well. Uncertainty, in any case, is inevitable in intellectual history.[6]

---

[2] Intellectual history involves two kinds of empirical work: research on actions and research on beliefs. A. Blau, "History of Political Thought as a Social Science," working paper, available on-line at http://www.socialsciences.manchester.ac.uk/disciplines/politics/about/themes/mancept/workingpapers/documents/AdrianBlauScienceofInterpretationdraft2March2010.pdf , 4 and "Uncertainty and the History of Ideas," *History and Theory*, 50 (2011), 359.

[3] Blau, "History of Political Thought as a Social Science," 2.

[4] M. Devitt, ''Scientific Realism,'' in F. Jackson and M. Smith (eds.), *The Oxford Handbook of Contemporary Philosophy* (Oxford: Oxford University Press, 2005), 768, 778.

[5] In fact, many intellectual historians proceed as if it would be obvious that underdetermination of interpretation by textual evidence is false.

[6] Blau "History of Political Thought as a Social Science," 10-11, "Uncertainty and the History of Ideas," 360-361 and "Anti-Strauss," *The Journal of Politics*, 7 (2012), 142, 147.

Underdetermination of theory by data has two interpretations: "weak" and "strong." The description I have given above is of weak underdetermination of theory by data. Weak underdetermination obtains when there are competing scientific explanations of the same data. However, this does not mean that the controversy cannot be decided one way or another. For instance, the rival theories can make different predictions on areas not yet investigated; seeking new data can thus help deciding between the competing explanations. In contrast, strong underdetermination of theory by data states that there are rival though empirically equivalent scientific theories.[7] In this case, no amount of "new" data can decide the controversy. Even "an observationally omniscient God who knew the observation states of the entire universe past, present, and future, would not be able to decide on that basis alone" between the rival explanations.[8]

Underdetermination of interpretation by evidence obtains in two sorts of cases. First, there are circumstances in which the textual evidence is very feeble and therefore scholars have to put forward more or less reasonable conjectures to make the most of it. In such cases, it is quite possible that two or more mutually inconsistent readings can be developed to account for the same evidence. So understood, underdetermination of interpretation by evidence can be assimilated to weak underdetermination of theory by data, although strictly speaking the two theses are not exactly the same. The evidence for

[7] See e.g. W. V. Quine, "On Empirically Equivalent Systems of the World," *Erkenntnis*, 9 (1975) for a classic account.

[8] W. H. Newton-Smith, "Underdetermination of Theory by Data," in W. H. Newton-Smith (ed.), *A Companion to the Philosophy of Science* (Oxford: Blackwell, 2000), 532. Although almost no one challenges the weak version, there is a heated dispute on the plausibility and even intelligibility of the strong version of the thesis among philosophers of science. See e.g. J. D. Norton, ''Must Evidence Underdetermine Theory?," in M. Carrier, D. Howard, and J. Kourany (eds.), *The Challenge of the Social and The Pressure of Practice: Science and Values Revisited* (Pittsburgh: University of Pittsburgh Press, 2008), 18-19, 27-28, 33-40 for a discussion.

grounding readings is finite, let alone limited, and so it can be the case that there are no novel areas on which to test the competing accounts.[9] Eventually, new textual evidence (e.g., a newly discovered manuscript) can be obtained and so used in a scholarly controversy. Carefully investigating this further evidence, when possible, is extremely important. However, it can simply happen that there is no such new evidence to explore. In turn, contextual findings can also be extremely useful in deciding a scholarly controversy, although it may also happen that the information we manage to obtain is not helpful for one reason or another.

There is a second form of underdetermination of interpretation that cannot be assimilated to either weak or strong underdetermination of theory by data. There are circumstances in which there is enough textual evidence to ground two or more mutually incompatible interpretations. Here the problem is not merely that the textual evidence is limited and/or obscure, thus allowing for rival interpretations. Rather, the issue is that the textual evidence itself is inconsistent, and so there is room for two or more incompatible interpretations. In other words, while in the first case of underdetermination there is *too little*, in the second case there is *too much* textual evidence, so to speak.

In this paper I am interested in analyzing the second form of underdetermination of interpretation. Taken at face value, it seems to involve the idea that there is no rational way to decide a given interpretive controversy. In case this form of underdetermination obtains, at most scholars may show that the evidence they are dealing with is contradictory, and

[9] Modern conceptual distinctions may be taken to enlarge the body of the whole evidence. However, they not enlarge the body of the textual evidence.

leave things there. This conclusion is too quick, though. In the next section, I will introduce a way to deal with underdetermination of interpretation by evidence.[10]

3. Crucial Arguments and Crucial Evidence

According to Quentin Skinner, assuming *a priori* that classic texts are fully coherent and articulated constitutes a dubious "mythology of coherence."[11] Skinner's complaint against this interpretative mistake has empirical and normative dimensions. Skinner's empirical claim is probably true. Underdetermination of interpretation assumes that texts can be incoherent, thus allowing for two or more mutually incompatible readings. However, the normative claim introduced by Skinner does not seem to be correct. The fact that texts might be incoherent does not entail that intellectual historians are not allowed to try to

[10] I would like to make a clarificatory remark before proceeding. The fact that I have identified two forms of underdetermination of interpretation by textual evidence might be taken to imply that I construe the relationship between evidence and interpretation in terms of logical entailment. Of course, this is not the case; rather, I think the opposite is true. To be sure, even when a text is fully coherent, we do not think that the evidence *entails* an interpretation. However, the consistent use of the terminology I am introducing would push me to construe such relation as one of *determination*. But this is innocuous, and so we are able to claim that the evidence merely *supports* a reading, even in those cases. The same relationship obtains in cases of underdetermination of interpretation, i.e., the evidence *supports* two or more inconsistent readings. Now, this does not attempt against introducing the very concept of underdetermination. The recognition of underdetermination of theory by data, and therefore of determination of theory by data, does not push philosophers of science to construe the relationship between data and theory in terms of logical entailment. The kind of support theories receive from data is independent of whether underdetermination obtains, and the same is true in the case of textual evidence and interpretation.

[11] Q. Skinner, *Visions of Politics. Volume I: Regarding Method* (Cambridge, UK: Cambridge University Press, 2002), 67-72. Admittedly, Skinner is mainly concerned with coherence across whole *œuvres*, i.e., series of texts written by the same author. But his denouncement also seems to involve a stronger thesis regarding coherence within single texts.

develop and ground one particular reading.[12] The evidence may be incoherent only in a superficial reading or when taken at face value or, perhaps, it may be incoherent only in a "local" but not in a "systemic" level. Among competing interpretations, one (and only one) might be correct, all things (including all the relevant textual evidence) considered. Of course, to claim that one interpretation among many plausible ones is accurate involves argument. The link between intellectual history and science can be helpful here.

Experimentation can help deciding between competing scientific theories. The philosophy of science has introduced a name for a particularly important kind of experiments: "crucial experiments." Crucial experiments provide decisive evidence for one among competing explanations.[13] We can introduce analogous ideas in the philosophy of intellectual history: "crucial arguments" or arguments that rely on "crucial evidence." The basic idea is that an argument is crucial if and only if it relies on especially important evidence—crucial evidence—to mediate in an interpretative controversy and provide decisive reasons to support one among different competing interpretations.[14]

---

[12] L. Venezia, "El cumplimiento de la obligación de obediencia al Leviatán: Hobbes, Skinner y la 'mitología de la coherencia,'" *Revista de Estudios Políticos*, 149 (2010), 160-167.

[13] See e.g. I. Lakatos, "The Role of Experiments in Science," *Studies in History and Philosophy of Science. Part A*, 4 (1974), 309. Robert Hooke coined the term *experimentum crucis* in his *Micrographia* of 1665. Issac Newton was the first natural philosopher who claimed having performed a crucial experiment in his studies on the nature of light of 1666-1672 that were publicly announced in the early 1670s. See S. Schaffer, "Glass Works: Newton's Prisms and the Uses of Experiment," in S. Schaffer (ed.), *The Uses of Experiment: Studies in the Natural Sciences* (Cambridge, UK: Cambridge University Press, 1989), 71-79 for a detailed account.

[14] Although I do not think this fact rules out the possibility of developing the analogy, it should be borne in mind that the two kinds of underdetermination at stake here are different. Crucial experiments provide decisive evidence for one among competing explanations when weak underdetermination of theory by data obtains. In turn, crucial arguments provide decisive evidence when the second form of underdetermination of interpretation by evidence (which cannot be assimilated to either weak or strong underdetermination of theory by evidence) obtains.

Some philosophers of science argue that crucial experiments are impossible. Famously, Pierre Duhem claims there are two reasons why this would be so. The first is that scientists do not test isolated hypothesis but rather a conjunction of hypothesis; in case the test fails, experiments do not establish which hypothesis should be rejected. The second problem is that the hypotheses tested by scientists do not exhaust the logical space of possible truths; there might be alternative explanations not yet considered.[15] Although these two arguments are sound, they do not undermine the possibility of developing crucial experiments. Here I follow Marcel Weber:

> Given what he sets out to prove, Duhem's arguments are impeccable. But note that Duhem is clearly thinking in terms of *deductive* inference. What he proves is that experiments conjoined with deductive logic, together, are unable to bring about a decision for one among a group of hypothesis. Of course, he is absolutely right about that. However, Duhem's arguments do not touch the possibility of *inductive* or *ampliative* inference enabling such a choice. An ampliative inference rule might very well be able to mark one hypothesis as the preferable one.[16]

Duhem's arguments do not undermine the possibility of developing crucial experiments, at least when construing them in Weber's revisionary fashion. In turn, Weber's point sheds light to the arguments employed in intellectual history. Although scholars use deductive logic to analyze conceptual claims, the empirical work done by intellectual historians fundamentally involves induction as well as inferences to the best

[15] P. Duhem, *The Aim and Structure of Physical Theory*, P. P. Wiener (trans.) (Princeton: Princeton University Press, 1954), 183-200. See also M. Weber, "The Crux of Crucial Experiments: Duhem's Problems and Inference to the Best Explanation," *British Journal for the Philosophy of Science*, 60 (2009), 21-22.
[16] Weber, "The Crux of Crucial Experiments," 22-23.

explanation.[17] For instance, intellectual historians use the latter to recover the author's intended meanings, i.e., what they understood by their own words, statements, passages, or texts.[18] In turn, scholars may also use inferences to the best explanation in cases in which there is conflicting evidence for different interpretations. Even in cases in which the second form of underdetermination obtains, there might still be a reading that gives the best interpretation of the whole evidence, taking into account the evidence for the contrary view.[19] In such circumstances, we would have strong reasons to support a particular interpretation instead of the alternative account.[20] The intuitive point here is that not all the evidence has the same "weight." Here is where the idea of crucial evidence is relevant. Crucial evidence is especially weighty evidence and so it allows scholars to claim that, despite the evidence to the contrary view, an alternative reading is to be preferred because it accommodates such evidence and therefore it is the best interpretation of the whole *corpus*.

The analogy between science and intellectual history I have been developing so far has a major shortcoming, though. In the case of crucial experiments, the data is not part of the empirical theory that explains it.[21] In the case of crucial arguments, however, the evidence is part of one of the competing readings, i.e., it is not new evidence.[22] This introduces an obvious worry. How can crucial evidence legitimately mediate in a scholarly

[17] Blau, "History of Political Thought as a Social Science," 7-8.
[18] Blau, "History of Political Thought as a Social Science," 8.
[19] Cf. Weber, "The Crux of Crucial Experiments," 23.
[20] This does not entail that the arguments are definitive, though. Without a doubt, plausible arguments to challenge the putative crucial character of the textual evidence may be developed. The introduction of the concepts of crucial arguments and crucial evidence in intellectual history is compatible with the thesis that states that the arguments put forward in the field are defeasible. See e.g. Skinner, *Visions of Politics I*, 121.
[21] Of course, this does not mean that the data is necessarily neutral. The data is always "theory-laden" and so, in a sense, it is related to the empirical theory that explains it.
[22] Even so, greater attention or a different perspective to a certain passage can be new.

controversy, instead of merely being provided as further textual evidence to support a particular reading?

We cannot completely overcome this difficulty. Each alleged piece of crucial textual evidence is always going to be part of one of the competing interpretations. However, we can minimize the risk of upgrading any evidence to the special status of crucial evidence and so of deciding scholarly controversies more or less by *fiat*. Besides supporting a particular interpretation in a given scholarly controversy, further conditions should obtain to legitimately claiming that a given piece of evidence is crucial. I introduce (b), (c) and (d) to the straightforward condition (a):

(a) The ideas put forward in the passage must clearly support a particular reading and so help deciding the controversy at stake.

(b) The interpretation of the passage must be relatively non-contested.

(c) The ideas introduced in the passage must play a key role in the author's broad theory.

(d) The ideas put forward in the passage must be (somehow) connected to other important theses of the broad theory.

(a) is the very minimal condition to consider crucial a given piece of evidence. To be sure, the passage in question should clearly favor one of the different competing interpretations; otherwise, it will not serve to decide any controversy. Condition (b) allows one to claim that one's own reading relies on relatively firm evidence. Condition (c) minimizes the risk of resolving the controversy by appealing to marginal ideas. If the ideas put forward in the passage are central rather than marginal, it seems more plausible to rely on them to decide among competing interpretations. Finally, condition (d) guarantees (at least to a certain extent) that the theses put forward in the passage are not *ad hoc*. If that

were the case, the whole reading might end up having this property too. If the ideas put forward in the passage are linked to other important theses, however, the risk of developing *ad hoc* interpretations diminishes. Thus, conditions (a)-(d) allow discarding evidence that is clearly non-crucial, although it may exist evidence that satisfies conditions (a)-(d) but that it may still be non-crucial. In any event, scholars have to exercise their judgment and evaluate whether the evidence is really crucial or not. There is no algorithm to decide this kind of issues.

There is a complication to be taken into account. Conditions (c) and (d) mention the author's "broad theory" as well as the "important theses" of her account. Since texts are not self-interpreting, it is quite possible that disagreement on these issues obtains too. Accordingly, to rely on these points would merely seem to move the controversy one step further instead of resolving it.[23] Although this point seems correct, we should not exaggerate it. At least some texts have central ideas, points and theses, and it seems plausible to consider that we can agree on what they are. An example can illustrate this issue. I imagine that Locke's discussion of "prerogative" in Chapter XIV of the *Second Treatise of Government* can legitimately produce some controversy on *how* liberal his whole political philosophy is. However, this does not seem enough to put pressure on whether Locke's theory of limited government *is* liberal. There are central theses in the *Second Treatise* to ground this idea. To be sure, Locke's thesis that we do not have a right to enslave ourselves is central. Locke relies on this claim to argue that it is not possible to owe a duty of obedience to an absolute government—a liberal position.[24]

[23] Blau, "History of Political Thought as a Social Science," 15-16.
[24] J. Locke, *The Second Treatise of Government*, in J. Locke, *Two Treatises of Government*, Peter Laslett (ed.) (Cambridge, UK: Cambridge University Press, 1988), §§ 23, 135, 137, 149, 171, 172.

The best way of introducing the concepts of crucial argument and crucial evidence is giving an example and using it in an ongoing scholarly controversy. In what follows, I will argue that Hobbes's texts underdetermine interpretation in the sense that they provide conflicting evidence for two rival readings of his theory of contractual obligation. However, I will also show that there is a way of deciding the controversy. I will claim that the passage in which Hobbes analyzes contracts made under coercion can play the role of crucial evidence and so can be used to decide the issue at stake. Finally, I will rely on Hobbes's example of the promise to the thief to support the deontological interpretation. First of all, however, I will provide a preliminary analysis of the very idea of contractual obligation.

4. Deontological Contracts and Self-Interested Agreements

According to Hobbes, the authority of the state depends on a contractual arrangement taken one way or another by the citizens. Hobbes's account of contractual obligation—and therefore his account of the agreement that grounds political authority—introduces the idea that agents incur new obligations when they voluntarily alienate or give up a part of their natural right and so bind themselves to act in a particular fashion. In other words, the claim that making a promise or signing a contract generates a new obligation on the agent characterizes both Hobbes's theory of contractual obligation in general and his contractarian theory of political authority in particular. In the following sections, I will only analyze his broad account of contractual obligation. In this section, I provide the conceptual framework to pursue this task. In particular, I introduce an intuitive deontological analysis

of the concept of contractual obligation as well as the revisionary theory of self-interested agreements put forward by Jean Hampton to analyze Hobbes's putative account.[25]

Intuitively, contracts have a clear deontological flavor.[26] Contracts are taken to introduce genuine obligations whose normativity is independent of the agents' contingent desires or interests in discharging them. The reason we undertake contractual obligations is one thing; the reason we ought to fulfill them is quite another. The obligation to perform the contract does not cease once my reasons for entering the contract disappear. We typically undertake obligations because it is in our rational self-interest to do so.[27] For what it is worth, the economic analysis of law assumes this idea.[28] However, this does not entail that the reasons to fulfill one's word are necessarily of the same kind. In particular, the fact that we have prudential reasons to make contracts does not entail that the reasons we have to fulfill them are necessarily linked to the promotion of our rational self-interest. However, we normally think this is not especially relevant. We have to comply because we have agreed to do so, independently of whether complying best promotes our desires and interests. Contracts properly performed introduce second-order reasons which exclude (some) first-order reasons for action (such as desires, interests, preferences, etc.) that,

[25] Several interpreters claim that Hobbes adopts a prudential theory of contractual obligation. I will only deal with Hampton's account in the paper because it introduces a distinctive conceptual framework to analyze Hobbes's theory that allows one to clearly distinguish two competing accounts.

[26] For stylistic reasons, in what follows I will only talk of contracts. However, this should be taken as shorthand for promises and contracts.

[27] This is contingent, though, for one may miscalculate or be really short-sighted and so take serious imprudent decisions only to discover one's mistake after assuming the obligation.

[28] The economic analysis of law assumes that agents normally make agreements in the expectation to be better-off and so that voluntary agreements are typically Pareto-superior. In addition, this approach to the law introduces the normative thesis that the law should enforce contracts.

absent the agreement, would have been sufficient to justify one course of action.[29] The relevant issue is that we have given our word and so voluntarily obliged ourselves which, in normal cases, is sufficient to make us act.[30]

The deontological character of contractual obligation can be illustrated with a mundane example. Say we have signed a lease to live in an apartment. Before signing the contract we considered there were good prudential reasons for us to choose this particular place, and so we signed it. In fact, we typically engage in cost-benefit reasoning when taking decisions like this. So far, so good. Time to pay the first month's rent. I take it that we consider we have to fulfill the agreement and so pay the rent because we have committed ourselves to do that rather than because doing so is in our rational self-interest. Of course, we may have additional prudential reasons to act in such a way (e.g., the owner may kick us out of the apartment). But this does not entail that we have to pay the rent just because this is in our rational self-interest to so act. In fact, we still would have reason to fulfill our word even if rational self-interest would recommend acting otherwise (e.g., because the owner of the apartment happens to be an old lady that cannot force us to pay).

These considerations notwithstanding, several scholars argue that Hobbes's moral and political philosophy introduces an idiosyncratic theory of contractual obligation that

[29] H. L. A. Hart, *Essays on Bentham: Studies in Jurisprudence and Political Theory* (Oxford: Clarendon Press, 1982), 255, J. Finnis, *Natural Laws and Natural Rights* (Oxford: Clarendon Press, 1980), 308, J. Raz, "Promises and Obligations," in P. M. S. Hacker and J. Raz (eds.), *Law, Morality and Society: Essays in Honour of H. L. A. Hart*, (Oxford: Oxford University Press, 1977), 221-223, 227-228 and *Practical Reason and Norms* (New York: Oxford University Press, 1999), 39, 190. For the record, they all analyze promises instead of contracts. I think the same analysis applies in both cases.

[30] Of course, there are issues revolving around contracts that turn them invalid *ab initio*. For example, circumstances such as coercion may involve that contracts are not fully voluntary and so morally binding. Also, the content of such voluntary undertakings may turn them invalid *ab initio*. For instance, an agreement to commit murder does not generate any moral obligation whatsoever. It should be borne in mind, though, that Hobbes's analysis does not introduce either of these two invalidating conditions.

construes this notion in a completely different fashion. For one, Hampton argues that for Hobbes contracts introduce obligations only when it is prudentially rational to fulfill them:

> for Hobbes, self-interest explains not only why we should do what we ought to do but also when our obligations arising from the surrender of right in a contract cease […]. This means that, according to Hobbes, *contractual obligations exist only insofar as it is in our interest to perform them*.[31]

Hampton analyzes Hobbes's underlying account for this theory as follows:

> Hobbes defines two conditions that must be met in order for an obligation to exist: First, there must be a renunciation or transfer of a right to another; second, it must be in the interest of the renouncer or transferer to respect that renunciation or transfer. So Hobbes defines the nature and extent of our obligations such that our performance of them can never conflict with self-interest.[32]

Based on this account, Hampton claims that Hobbes is committed to the idea that contracts are binding insofar—and *only* insofar—it is prudentially rational to act on them.

Hampton acknowledges that the standard analysis of contracts introduces the idea that contracts have a deontological flavor. For this reason, she claims that Hobbes's contractarianism does not really involve genuine contracts but rather the highly idiosyncratic arrangements she dubs "self-interested agreements." Hampton claims that "In SI [self-interested] agreements, self-interested rational calculation, rather than the sense of 'duty' arising out of a promise or fear of a coercive power, is the motive for each person's

[31] J. Hampton, *Hobbes and the Social Contract Tradition* (Cambridge, UK: Cambridge University Press, 1986), 56.
[32] Hampton, *Hobbes and the Social Contract Tradition*, 56.

performance of the act agreed on."[33] Accordingly, self-interested agreements "differ from contracts in being coordination of intentions to act that are kept by both parties *solely for self-interested reasons*, whereas contracts are trades of *promises* that introduce moral incentives that either *supplement* or *replace* each party's self-interested motivations."[34] Thus, contrary to genuine contracts, self-interested agreements do not possess a deontological character: "SI agreements fundamentally differ from contracts in that the 'benefits of the bargain' are sufficient to motivate the parties to perform the actions agreed on."[35]

5. Hobbes on Contractual Obligation

We can now proceed to examine the textual evidence. Taken at face value, Hobbes's various remarks underdetermine a deontological as well as a prudential interpretation of his theory of contractual obligation in the sense that they provide conflicting evidence for the two rival readings. In this section, I give the textual evidence. I start with the prudential account.[36]

[33] Hampton, *Hobbes and the Social Contract Tradition*, 139. For the record, Hampton's conflation of genuine moral duty with fear of threats of punishment is wrong. The two are analytically distinct notions. In fact, the latter should be understood along prudential lines.
[34] Hampton, *Hobbes and the Social Contract Tradition*, 145-147.
[35] Hampton, *Hobbes and the Social Contract Tradition*, 142; see 138-147 for Hampton's full analysis of self-interested agreements.
[36] Unless indicated, in what follows I quote *Leviathan: with selected variants from the Latin edition of 1668*, E. Curley (ed.) (Indianapolis: Hackett, 1994) in the body of the paper and I refer in the footnotes to corresponding or related passages of *The Elements of Law, Natural and Politic*, F. Tönnies (ed.) (London: Frank Cass & Co., 1969) and *De Cive*, translated as *On the Citizen*, R. Tuck and M. Silversthorne (eds.) (Cambridge, UK: Cambridge University Press. 1998.

Hobbes writes that after having alienated or granted away rights agents become "obliged" or "bound." Such voluntary actions introduce a contractual obligation[37] "not to hinder those to whom such right is granted or abandoned from the benefit of it."[38] The crucial feature of Hobbes's analysis is that Hobbes then argues that such bonds "have their strength, not from their own nature (for noting is more easily broken than a man's word) but from fear of some evil consequence upon the rupture."[39] Hobbes's description of the way agents become bound to respect the beneficiaries of their renunciation of rights is prudential. They have a contractual obligation to respect them because of fear of the consequences of breaking such undertakings.

Hobbes's analysis the third law of nature—that prescribes "*that men perform their covenants made*"[40]—also introduces prudential elements. In turn, he claims that

> in this law of nature consisteth the fountain and original of JUSTICE. For where no covenant hath preceded, there hath no right been transferred, and every man has right to everything; and consequently, no action can be unjust. But when a covenant is made, then to break it is *unjust*; and the definition of INJUSTICE is no other than *the not performance of covenant*. And whatsoever is not unjust, is *just*.[41]

[37] Strictly speaking, Hobbes here talks about "duty" rather than "obligation." Contrary to the modern use, he does not make differences between these two kinds of requirements.
[38] *Leviathan* XIV.7.
[39] *Leviathan* XIV.7.
[40] *Leviathan* XV.1. Cf. *The Elements of Law* I.XVI.1 and *De Cive* III.1.
[41] *Leviathan* XV.2. Cf. *The Elements of Law* I.XVI.2 and *De Cive* III.3. Notice that in *The Elements of Law* I.XVI.2 Hobbes says that the breach of the duty to keep one's own word constitutes "injury."

The third law of nature does not say that one should keep one's word insofar as one has self-interested reasons to so act.[42] However, Hobbes's further discussion points in a prudential direction. In particular, he makes important claims in this sense in his further argument with the "fool" who puts pressure on the rationality of contract-keeping.

Hobbes claims that the fool does not question the existence of contracts nor does he question the injustice of breaking them. He writes that the fool "does not […] deny that there be covenants, and that they are sometimes broken, sometimes kept, and that such breach of them may be called injustice, and the observance of them justice."[43] However, the fool says injustice may be consistent with "that reason which dictateth to every man his own good." The fool thus claims:

> "The kingdom of God is gotten by violence; but what if it could be gotten by unjust violence? were it against [OL: right] reason so to get it, when it is impossible to receive hurt by it [OL: but only by supreme good]? and if it be not against reason, it is not against justice; or else justice is not to be approved for good."[44]

According to the fool, if acting unjustly is beneficial to the agents, then either injustice is compatible with reason or justice is not necessarily good or desirable. Hobbes replies to this point attempting to show that keeping one's word is prudentially warranted in the commonwealth. Hobbes gives three different arguments to ground the point that justice is compatible with reason because *ex ante* rational self-interest recommends fulfilling one's

[42] Hobbes does not talk here of contracts but of covenants. The distinction between these two concepts is irrelevant here, although it will be relevant bellow.
[43] *Leviathan* XV.4.
[44] *Leviathan* XV.4.

own word.[45] This way, Hobbes's reply to the fool provides important textual resources to ground a prudential account of contractual obligation.[46]

Although the prudential reading is well-established in the *corpus*, there is also important textual evidence to construe Hobbes's theory of contractual obligation in a deontological fashion. I now provide the different passages that point in this direction. I start with Hobbes's account of covenants.

Hobbes's account of covenants introduces a strict deontological theory of contractual obligation. Hobbes characterizes a covenant as a contractual agreement in which at least one party does not perform his or her part of the bargain immediately but rather promises to do so in the future. Hobbes's account thus introduces the claim that such promises create moral obligations to be later discharged. Although Hobbes does not make

[45] See *Leviathan* XV.5-7. Reasons of space do not allow me to engage with the arguments put forward by Hobbes to block the fool's challenge. They are well-known, though.

[46] Although the scholarly literature on Hobbes's reply to the fool is enormous, the crucial works to mention here are the ones that make use Hampton's conceptual machinery to interpret the passage. Of course, Hampton herself argues that that Hobbes's reply to the fool is articulated in terms of self-interested agreements rather than deontological contracts. In addition, D. Gauthier, "Taming Leviathan," *Philosophy and Public Affairs*, 16 (1987), 295 develops a related reading of the passage, although it should be borne in mind that Gauthier also makes it clear that he does not share Hampton's interpretation of Hobbes. See D. Gauthier, *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes* (Oxford: Clarendon Press, 1969), 60-61, "Taming Leviathan," 295-296 and "Hobbes's Social Contract," in G. A. J. Rogers and A. Ryan (eds.), *Perspectives on Thomas Hobbes* (Oxford: Clarendon Press, 1988), 134-138 for his own interpretation. Gauthier claims that Hobbes's formal account of obligation is not necessarily related to prudence, i.e., he grants that agents may possess contractual obligations independently of whether or not prudence recommends complying with them. But this is not the end of the story for Gauthier. He reformulates Hobbes's account in such a way that these contracts are not "reasonable," where a contract is reasonable if and only if it is rational to enter *and* to keep it. This allows him to argue that "Hobbes needs no longer reply that if a covenant is undertaken to secure peace and preservation it must be reasonable to fulfill it. Instead he can admit that, in the state of nature, it may not be reasonable to fulfill such a covenant—and therefore some power is needed to *make* it reasonable to fulfill it" (*The Logic of Leviathan*, 89; see 76-89 for his full analysis).

this point really plain in his analyses of covenants in *The Elements of Law* and *Leviathan*,[47] he does make it crystal clear in the account he introduces in *De Cive*:[48]

> An *agreement* made by a party who is trusted with a party who has already performed, even if the promise is made in words referring to the future, is no less a transfer of a *right* at a future time than if it had been made in words referring to the present or the past. For performance is the most evident sign that one who has performed his part understood the words of the other party (the trusted party) as expressing an intention to perform at a specified time. By that sign also the trusted party knew that he was so understood; and because he did not correct it, he intended it to be so taken. *Promises* therefore which are made in return for *good* received (such promises too are *agreements*) are signs of will, that is […] signs of the last act of deliberation by which the liberty not to perform is lost; consequently they are obligatory; for obligation begins where liberty ends.[49]

Hobbes's analysis of covenants is straightforwardly deontological. As he writes, covenants put an end to liberty and create an obligation to act in a particular fashion. After promising to act in a certain way, agents bind themselves and so forfeit the liberty to act against what they promised. For the agents the question now is settled and so they do not have the liberty to deliberate about it but rather they have a moral obligation to keep their word which settles the issue.

---

[47] See *The Elements of Law* I.XV.9 and *Leviathan* XIV.11.

[48] For this reason, some writers claim that Hobbes modified his account of contractual obligation between 1647 and 1651. See e.g. D. Eggers, "Liberty and Contractual Obligation in Hobbes," *Hobbes Studies*, 22 (2009), 76-77, 93-96. I do not think the whole evidence allows making such a strong claim. In my view, at most, it suggests a change of emphasis.

[49] *De Cive* II.10.

Hobbes also introduces a deontological account of contractual obligation in his analysis of oaths. He writes that "*swearing*, or OATH, is a *form of speech, added to a promise, by which he that promiseth signifieth that unless he perform, he renounceth the mercy of his God, or calleth to him for vengeance on himself*."[50] In addition, he states that

> It appears also that oath adds nothing to the obligation. For a covenant, if lawful, binds in the sight of God without the oath as much as with it; if unlawful, bindeth not at all, though it be confirmed with an oath.[51]

When analyzing oaths, Hobbes discusses whether an oath attached to a promise adds something to the obligation to keep it. Hobbes claims that voluntary undertakings have obligatory force in themselves and thus that oaths add nothing to this pre-existing obligation. While contracts are genuinely normative, oaths do not add anything. At most, they provoke negative consequences in the case of non-compliance.[52] This way, Hobbes's analysis introduces the idea that contractual obligation is *not* to be associated with negative consequences in the case of non-compliance. Instead, his remarks involve the thesis that contractual obligations are normative in themselves.[53]

Finally, Hobbes's discussion of the particular contracts signed by soldiers introduces further evidence to claim that he holds the view according to which binding contractual obligations are not necessarily prudential but that rather that they introduce categorical moral requirements. Hobbes argues that "he that enrolleth himself as a soldier, or taketh imprest money, taketh away the excuse of a timorous nature, and is obliged, not

[50] *Leviathan* XIV.31. Cf. *The Elements of Law* I.XV.15 and *De Cive* II.20.
[51] *Leviathan* XIV.33. Cf. *The Elements of Law* I.XV.17 and *De Cive* II.22.
[52] This point is present in the analyses of oaths in *The Elements of Law* I.XV.17 and *De Cive* II.22 but is absent in the correlative passage in *Leviathan* XIV.33.
[53] See also Eggers, "Liberty and Contractual Obligation in Hobbes," 77 for a related analysis.

only to go to the battle, but also not to run from it without his captain's leave."[54] Hobbes's discussion of the particular contracts signed by soldiers seems to involve the idea that contracts impose more stringent obligations on those that perform them than the moral requirements imposed on all persons by the laws of nature. Although in this paper I have not discussed the kind of duties introduced by the laws of nature, for the sake of the argument I will grant it here that they are self-interested maxims that introduce prudential requirements.[55] Even so, Hobbes's analysis of the contracts signed by soldiers shows that such arrangements introduce moral obligations that have to be discharged even if agents do not maximally promote their desires and interests when so acting. To be sure, prudence does not recommend fulfilling such contractual obligations. To begin with, people endanger their health and survival when engaging in combat. In fact, it seems very difficult to claim that participating in warfare may be advantageous to the agents involved. On the contrary, when evaluating the issue from a prudential perspective, it seems that the actual behavior recommended by rational self-interest is contrary to discharging this contractual obligation. Instead, prudence seems to recommend other ways of acting (e.g., refusing to participate in combat even when that may put oneself in jail, changing sides when the probabilities of winning were low, seeking asylum in neutral states, etc.). However, this does not matter for Hobbes. Rather, he claims that soldiers must keep their word and fight even if that entails risking their lives. In fact, Hobbes's account seems to involve the idea that enlisted soldiers give up their right of self-defense in the sense that they have to stay and fight in combat,

[54] *Leviathan* XXI.16.

[55] I do not think that the laws of nature uniquely impose this kind of requirements but also genuinely moral duties. Reasons of space do not me allow to develop my own reading here.

even when doing so involves their own death.[56] Thus, Hobbes's analysis of the contracts signed by soldiers introduces further evidence to ground the reading according to which contracts introduce deontological moral requirements to be discharged even when there are alternative forms of behavior that might best promote one's desires and interests.[57]

There is conflicting textual evidence to interpret whether Hobbes's account of contractual obligation introduces the idea that contractual obligation is necessarily linked to rational self-interest or whether he analyses contractual obligation in deontological terms. On the one hand, Hobbes's description of the bonds created when having abandoned or granted away rights is prudential. Also, the reply to the fool supports the prudential reading. On the other hand, however, Hobbes's account of covenants, oaths and the contracts signed by soldiers supports the deontological interpretation. Hobbes's texts thus underdetermine interpretation, although it seems that the bulk of the evidence analyzed so far supports a deontological reading of his theory of contractual obligation. Some people may think that there is no rational way to settle the controversy at stake and so they may prefer to leave things as they are. Instead, I will argue that there is further evidence—crucial evidence—to decide the issue.[58]

---

[56] The issue of whether one can give up one's own right of self-defense in general is a matter of scholarly controversy. Although I prefer not taking sides in the debate, it seems clear that the contracts signed by soldiers involve alienating such a right.

[57] See also D. Baumgold, "Subjects and Soldiers: Hobbes on Military Service," *History of Political Thought*, 4 (1983), 59-62 and S. Sreedhar, *Hobbes on Resistance: Defying the Leviathan* (New York: Cambridge University Press, 2010), 38-40, 84-87 for further analysis of contracts signed by soldiers.

[58] Hobbes's point that violation of contracts involves a form of self-contradiction in *The Elements of Law* I.XVI.2, *De Cive* III.3 and *Leviathan* XIV.7 may also be taken as providing important evidence for the claim that he develops a deontological (in particular, Kantian) theory of moral obligation. In turn, Hobbes's analogy may be taken as providing further evidence for the deontological reading of his theory of contractual obligation. However, Eggers, "Liberty and Contractual Obligation in Hobbes," 78-79 shows that the

6. Contracts Made under Coercion

Famously, Hobbes writes that

> Covenants entered into by fear, in the condition of mere nature, are obligatory. For example, if I covenant to pay ransom, or service, for my life, to an enemy, I am bound to it. For it is a contract wherein one receiveth the benefit of life; the other is to receive money, or service, for it; and consequently, where no other law (as in the condition of mere nature) forbiddeth the performance, the covenant is valid.[59]

Hobbes does not argue that promises and contracts entered into by fear are binding only in the state of nature. On the contrary, Hobbes explicitly extends this principle to civil society. He says that "even in commonwealths, if I be forced to redeem my self from a thief by promising him money, I am bound to pay it, till the civil law discharge me."[60]

Hobbes introduces these examples to ground his thesis that contracts entered into by fear are valid and therefore morally binding. This passage constitutes crucial piece of evidence to ground the deontological interpretation of his theory of contractual obligation.

The passage fulfills conditions (a)-(d) for crucial evidence. As I will show in the following section, the passage clearly supports the deontological reading of Hobbes's theory of contractual obligation, thus satisfying conditions (a)-(b) that states that the passage must clearly support one particular reading and so help deciding the controversy at stake and, also, that the interpretation of the very passage should be non-contested.

---

passages involve serious problems and so that we should not put much interpretive weight on them.

[59] *Leviathan* XIV.27. Cf. *The Elements of Law* I.XV.13 and *De Cive* II.16.

[60] *Leviathan* XIV.27. Of course, Hobbes's analysis does not entail the idea of the robber having a right to assault the victim.

The passage also satisfies the condition (c) that states that the thesis introduced must play a major role in the author's whole theory. The passage's underlying theory plays a key role in Hobbes's broad political philosophy, notably in his account of sovereignty. Hobbes's contractarianism relies on the idea that contracts performed under coercion are morally binding. Hobbes argues that political authority and obligation is grounded in an agreement made by subjects, whether it be attained by "acquisition" or by "institution."[61] Hobbes analyses both sovereignty by acquisition and sovereignty by institution as the result of agreements performed for the same kind of reason, namely, fear of death.[62] Thus, Hobbes's discussion of contracts made under coercion plays a major role in Hobbes's account of political obligation, arguably the main subject of Hobbes's political philosophy.[63]

Finally, the passage also satisfies condition (d) that states that the ideas put forward in the passage must be connected to other important theses of the author. The thesis introduced in the passage is not unrelated to the rest of Hobbes's theoretical commitments. On the contrary, the theory Hobbes puts forward when discussing contracts made under coercion is closely linked to other substantial theses, especially his account of voluntariness. Hobbes writes that

---

[61] Hobbes introduces this terminology in *Leviathan* XX.1-2.

[62] However, Hobbes seems to be conflating two disparate situations here. In one case, the fear is the product of another agent's deliberate threat, while in the other case the fear is merely of the consequences of the joint actions of other persons that do not directly introduce threats. See A. J. Simmons, "Theories of the State," in D. Rutherford (ed.), *The Cambridge Companion to Early Modern Philosophy* (Cambridge, UK: Cambridge University Press, 2006), 273 n. 14.

[63] See e.g. Gauthier, *The Logic of Leviathan*, 41, Q. Skinner, *Reason and Rhetoric in the Philosophy of Hobbes* (Cambridge, UK: Cambridge University Press, 1996), 389 and H. Warrender, *The Political Philosophy of Hobbes: His Theory of Obligation* (Oxford: Clarendon Press, 1957), 326 for different claims in this vein.

> In deliberation, the last appetite or aversion immediately adhering to the action, or to the omission thereof, is that we call the WILL, the act (not the faculty) of *willing*. […] a *voluntary act* is that which proceedeth from the will, and no other. […] Will therefore is the last appetite in deliberating.[64]

Hobbes's analysis introduces the idea that "not only actions that have their beginning from covetousness, ambition, lust, or other appetites to the thing propounded, but also those that have their beginning from aversion or fear of those consequences that follow the omission are *voluntary actions*."[65] Hobbes's account of why contracts made under coercion are morally binding is linked to his analysis of voluntariness. Hobbes analyses the will as the last desire of agents after deliberating about what to do. This implies the idea that contracts made under coercion are completely voluntary and intentional. Coerced contracts are entered into by fear and, insofar as this does not prevent deliberation, for Hobbes they are fully voluntary and thus morally binding.

The passage in which Hobbes analyses contracts made under coercion is a piece of crucial evidence. In the following section, I will argue that it allows claiming that the best explanation of the whole evidence—including the evidence for the alternative reading—is that Hobbes's account of contractual obligation is deontological.[66]

---

[64] *Leviathan* VI.53. Cf. *The Elements of Law* I.XII.2, I.XV.7.
[65] *Leviathan* VI.54.
[66] The passage is also important for other reasons. Skinner argues that Hobbes's theory is related to specific historical circumstances, particularly to the events that took place in England between January and August of 1649. Moreover, Skinner claims that the very intelligibility of Hobbes's account necessarily requires a proper understanding of the historical events to which it is related. See Q. Skinner and Y-C Zarka, *Hobbes: The Amsterdam Debate*, Hans Blom (ed.) (Hildesheim: Georg Olms Verlag, 2001), 28. See also Q. Skinner, *Visions of Politics. Volume III: Hobbes and Civil Science* (Cambridge, UK: Cambridge University Press, 2002), 303-307 for an analysis of the relation between Hobbes's account and the *de facto* theories developed during the "engagement controversy," i.e., the debate regarding whether those citizens who had previously taken an

7. Contractual Obligation and Reason

According to the account put forward in Hobbes's discussion of contracts made under coercion, the feature that really matters in evaluating the validity of a contract is that the contract was undertaken. The reason for which agents incur such obligations (crucially, whether or not they are incurred out of coercion) does not play any role in his account. Also, whether agents would best satisfy their desires and interests when acting on their word do not seem to play a significant role for him either. In fact, Hobbes's analysis of contracts made under coercion shows that for him agents should fulfill their contractual obligations even though doing so may not be advantageous for them.[67] The fact that fulfilling a contract made under coercion (or, for what matters, in normal circumstances) may not be prudent and may actually be irrational from a prudential point of view does not render such obligation invalid. After making the promise or agreement, they are "obliged" and "bound." And these are moral notions, as Hobbes makes it clear when stating that

> prisoners of war, if trusted with the payment of their ransom, are obliged to pay it; and if a weaker prince make a disadvantageous peace with a stronger, for fear, he is bound to keep it, unless […] there ariseth now new and just cause of fear, to renew the war.[68]

This obligation then is not prudential, but of a moral kind. For Hobbes, the performative action of signing a contract morally binds the agents, and so they now have a moral reason

---

oath to protect the life and person of the king could be justified in taking a new oath of allegiance to the Commonwealth. But see also Skinner, *Visions of Politics III*, 232 for a much less emphatic interpretation.

[67] Here I just assume that the various invalidating conditions introduced by Hobbes have not been met.

[68] *Leviathan* XIV.27. Cf. *The Elements of Law* I.XV.13 and *De Cive* II.16.

to discharge such obligations that is sufficient to motivate them to act even in those cases in which they do not have prudential reasons to so act.

Let us investigate this point carefully. The first thing to notice is the following. The fact of having made a contract for self-interested reasons does not say much about the kind of reasons one may have for keeping one's word. The example of the promise to the thief perfectly illustrates this point, and in fact shows that *ex post* agents may not have prudential reasons to discharge contractual obligations even though they would have been undertaken by these reasons. Even though the victim may have promised for prudential or self-interested reasons—which of course might have been the case, since she is being coerced by the thief for this to occur—once released she would not have prudential reasons to keep her word. Indeed, rational self-interest surely would recommend breaking the promise. For instance, she may consider going to the police (or calling some friends, or contacting the mafia, etc.) or simply leaving things as they are. But it is pretty obvious that she would not really have prudential reasons to keep her promise.[69]

In fact, when evaluating the different alternatives open to the agent from a strict prudential or self-interested point of view, the idea of going back to the thief to fulfill her promise seems to constitute a quintessentially irrational action. *Ceteris paribus*, facing the thief again would not only reduce the amount of money of the now "ex-victim," but would also expose her to be coerced one-more-time, which would put her in the (painful) situation of having to promise again in order to be released for the second time, and so to renew the cycle promise-release-compliance once again. Keeping her word constitutes a non-compensated cost for the ex-victim once she has been released by the thief with this very

[69] See also D. van Mill, *Liberty, Rationality, and Agency in Hobbes's* Leviathan (Albany: State University of New York Press, 2001), 132-133.

purpose as the sole objective. Accordingly, she would not have prudential reasons to act in such a way.

The prudential irrationality of complying with the promise raises an important worry. Were this situation to actually obtain, it seems that Hobbes would say that the very validity of the promise to the thief would be put into question in the first place. Using the terminology introduced in *Leviathan* XIV.18,[70] Hobbes should have to say that, in case the thief releases the victim, he would be "betraying himself to his enemy" insofar as the thief would have "no assurance" that the victim would comply with her word. In fact, he would have the "reasonable suspicion" that the victim would never keep her word even if the thief would release her only after promising to do so.[71] For this reason, the validity of the promise would be put into question and thus the obligation to fulfill the promise would vanish. The fact that the thief does not seem to have reason to trust the victim would strike against the validity of the very promise of paying money. However, Hobbes does *not* consider the promise to the thief void, *nor* does he consider fulfillment of the promise irrational.

The apparent anomaly lies in the fact that in the above analysis a key step is missing: the very act of making the promise to the thief. In effect, the interaction sequence between the victim and the thief does not involve two but three steps. The first move is made by the victim; she would have to decide whether to make the promise or not. In case the victim would have promised to pay the thief a certain amount, second, the liberation of the victim would occur or not. In case the thief would have liberated the victim, third, the

[70] Cf. *The Elements of Law* I.XV.10 and *De Cive* II.11, XIII.7.
[71] By the way, contrary to what is suggested here, the interaction between the thief and the victim takes place in civil society. However, they cannot trust that the sovereign will oblige the victim to fulfill her promise—the thief's act is out of law and so he denies his authority.

now ex-victim would have to comply with her promise or not. In this scenario, it would be the ex-victim the one who would be fulfilling her part when first making the promise to the thief; but she would not have trouble trusting that the thief would release her. For this reason, she would not be "betraying herself to her enemy" when making the promise insofar as she knows that the thief does have the intention to release her after she promised—the thief is threatening the victim to force her to act in that way. In turn, from the point of view of the thief, it suffices that the probability of compliance of the ex-victim be positive for him to act rationally when liberating her (after all, by construction, in Hobbes's example the only way of obtaining the money is by releasing the victim). It seems fair to assume that the thief may consider that the probability of the ex-victim keeping her word to be greater than zero, which thus entails that the thief would not be "betraying himself to his enemy" when releasing the victim. Accordingly, he would not be acting irrationally when releasing the victim in order to let her get the money.[72]

Although this analysis does not have the problem mentioned above, it seems that it nevertheless possesses a fatal shortcoming. It looks as if it would be rational for the victim to promise paying some money in exchange of being released by the thief. In fact, it seems rational to promise almost anything to the thief, and of course promising him a good amount satisfies this condition. However, this does not seem really possible if we assume

[72] Strictly speaking, even if the probability of the victim fulfilling her word were zero, the act of liberating her by the thief would not be irrational. In this case, the thief would have to adopt a random procedure to make the decision of whether to liberate her (e.g., flipping a coin). And there are other elements that may push the thief to release the victim as well. For instance, keeping the victim captive involves a cost for the thief (he needs a place to keep her, he has to feed her, etc.) and in addition he also exposes himself to being caught and so to being punished as a kidnapper. In case of liberating the victim, he would free himself of such costs.

that by promising the agent incurs a commitment.[73] The victim would not be able rationally to promise to pay the thief money in order for him to release her insofar as the situation faced by the victim looks structurally analogous to Kavka's "toxin puzzle."[74]

This puzzle goes as follows. A millionaire, who also is an excellent intention-guesser, offers an agent a million dollars if she forms at midnight the intention of drinking a toxin the day that follows. The toxin will make the agent sick, but it will not kill her, and in addition it is stipulated that the toxin does not possess side effects. *Prima facie*, the rational decision for the agent to make seems obvious: to form the intention of drinking the toxin. This will allow her to earn a million dollars at a reasonable cost. But part of the deal is that the agent does not have to effectively drink the toxin; to get the money she just has to form the intention of drinking it. And here is the paradoxical result of the toxin puzzle: insofar as the money will be in the agent's pocket the day that follows, drinking the toxin does not carry any economic benefit for her; rather, it involves a non-compensated cost, which thus gives her reason not to drink it. However, of course, the agent knows this before forming the intention place, which in turn makes her unable to form the intention to drink the toxin in the first place. In case of forming the intention, she would be committing herself to something irrational. Thus, the agent cannot get the million dollars she is willing to get in

[73] I take it that this is reasonable. After A has sincerely promised (not) to φ, it seems reasonable to assume that A has the intention of (not) φ-ing. See *The Elements of Law* I.XII.9, I.XIV.6.

[74] C. Finkelstein, "A Puzzle about Hobbes on Self-Defense," *Pacific Philosophical Quarterly*, 82 (2001), 345-348 introduces the toxin puzzle to analyze Hobbes's argument for the right of self-defense in civil society. I decided to use it to analyze the promise to the thief after reading her analysis, although her conclusions are very different from my own conclusions.

exchange for drinking the toxin. Even though the deal seems reasonable, she cannot make it, at pain of acting irrationally.[75]

Analogously to Kavka's toxin puzzle, it seems that in Hobbes's example the victim would not be able to form the intention to pay the thief once released, even if so acting would benefit her at the moment of making the promise. The crucial point is that the action of fulfilling her word would clearly be irrational from a prudential perspective. In such context, she would not have to bother with being liberated—at the time she would be free—but would have to consider only her money. As I have said above, keeping her word would be prudentially irrational. But, as in the toxin puzzle, this entails that she would not be able to form the intention of keeping her word in the first place and, so, she would not be able to make the promise to the thief for him to release her. In case of so acting, she would be forming the intention of doing something she would not have reason to do—and a rational person cannot act in such a fashion.

Accordingly, it looks like the very act of making the promise cannot take place and so the situation of paying the thief cannot happen either. But Hobbes does claim that the ex-victim would have to fulfill her word insofar as her life would not be in real danger and there is no civil law that prohibits the deal, at least till the sovereign liberates the ex-victim of her contractual obligation. Hobbes thus assumes that making the promise *is* rational, besides the point that the victim would possess the genuine obligation to fulfill her word. For this reason, there seems to be something missing in the analysis of the promise to the thief in the context of the above rational-choice framework—something that makes

[75] G. S. Kavka, "The Toxin Puzzle," *Analysis*, 43 (1983).

Hobbes's claim unacceptable, unless we play a new move. As far as I can see, the only way of solving the riddle is to introduce non-prudential elements to analyze the case.[76]

The most reasonable way of explaining Hobbes's intuition that both making the promise and fulfilling the promise is rational is the following. Making the promise is rational for the victim insofar as it involves bypassing the toxin puzzle: even though it is not rational to form the intention of doing something that is irrational to do, it is rational to make a promise to do something that only *ex ante* is irrational to perform, insofar as the fact itself of making a valid promise changes the normative scenario, thus turning compliance rational rather than irrational. After promising, the victim would have a binding moral obligation to do something (paying the thief some amount of money) while she would not have such an obligation before promising. In turn, this obligation would give the ex-victim a new reason—a moral reason—which would make fulfilling the promise rational.

Let me now make a clarificatory point. To be sure, we do not think we would acquire any obligation when making a contract under coercion. Accordingly, we would not

[76] One may eventually try a move *à la* Gauthier to show that the victim would rationally form the intention to pay the thief and so that she would be able to promise acting in such a way. Gauthier writes that "I grant, of course, that drinking the toxin doesn't have best, or even good, consequences, so that I have no *outcome*-oriented reason to drink it. But drinking the toxin is part of the *best course of action*—in terms of its consequences—that I can embrace as a whole. For I do better to intend to drink the toxin, even at the cost of actually drinking it, than not to intend to drink the toxin. And although I should do better still to intend to drink the toxin but not drink it, I cannot embrace this as a single course of action" (D. Gauthier, "Rethinking the Toxin Puzzle," in J. L. Coleman and C. W. Morris (eds.), *Rational Commitment and Social Justice*: *Essays for Gregory Kavka* (Cambridge, UK: Cambridge University Press, 1998), 48). Of course, the scenario described by Gauthier is better all things considered. However, this does not show that his move is adequate. For one, D. Skyrms, *Evolution of the Social Contract* (Cambridge, UK: Cambridge University Press, 1996), 24 argues that rationality is "modular" in the sense that it involves the fact that when an agent faces a sequence of actions her plan of action must specify that each decision is rational in each and every option. Gauthier's argument is inadequate precisely because it violates this principle of rational choice. The "course of action" or plan introduced by Gauthier involves that when facing the choice node (to drink the toxin / not to drink the toxin) the agent has to do something irrational (drinking the toxin).

grant that fulfilling such contracts would be rational because we would have moral reasons to so act. But this would be the case because we think we would lack the new reason; our evaluation would be different in case we grant that contracts made under coercion create genuine obligations. For what it is worth, we think that it is rational to make and fulfill a contract that only *ex ante* is irrational to discharge if the contract is made in normal (i.e., non-coerced) circumstances. Thus, the solution in itself makes sense; our problem is Hobbes's claim that contracts made under coercion introduce genuine moral obligations and so create new reasons for action.

This analysis of Hobbes's example solves the problem of the rationality of both making and fulfilling the promise to the thief. The making of the promise would be motivated uniquely by prudential or self-interested considerations, while compliance would be motivated by moral reasons other than prudential considerations. In particular, the fact of having made a valid promise would have changed the normative scenario and thus would have introduced a new obligation. That would have provided the agent with a moral reason to ground the rationality of compliance.[77] In this framework, the victim would be able to make the promise because she knows that performing it would be rational to keep it once released, and she would also possess sufficient reason to comply. So analyzed, Hobbes's example shows that for him contracts introduce elements different than mere narrow rational self-interest; they create obligations that introduce moral reasons to act.

Once we have reached this point, it seems fair to argue that claiming that contractual obligations are valid only insofar as it is prudentially rational to discharge them does not do full justice to Hobbes's account. On the contrary, Hobbes believes that the mere fact of

[77] See also M. E. Bratman, "Toxin, Temptation, and the Stability of Intention," in Coleman and Morris (eds.), *Rational Commitment and Social Justice*, 63-64.

having concluded a contract is almost the only relevant element for the validity of the act at stake, thus creating the obligation to comply. Other considerations that one may consider relevant (such as the reasons that motivated the act, the content of the different speech acts, the personal cost involved in complying, the general consequences of compliance, etc.) play a marginal role in his account. Despite some evidence to the contrary theory, Hobbes's analysis of contracts made under coercion allows us to claim that he develops a genuinely deontological account of contractual obligation; he believes that agents do have to keep their word because they have voluntarily obliged themselves to so act. To be sure, his full theory of contractual obligation also introduces the idea that there are features that invalidate contracts. However, they are very few, and apply mainly in the state of nature. Crucially, the contingent feature that compliance fails to maximally fulfill the agents' desires or preferences is not among them.

8. Final Remarks

In this paper I have identified a distinctive form of underdetermination: a particularly pressing form of underdetermination of interpretation by textual evidence that obtains when the evidence itself is contradictory and so it supports different readings. In turn, I have introduced the concepts of crucial argument and crucial evidence to deal with this issue. Finally, I have grounded the deontological interpretation of Hobbes's theory of contractual obligation in a particular piece of crucial evidence—his description of contracts made under coercion.

I do not claim that the deontological reading of Hobbes's theory is definitive. Critics can put pressure on whether the passage I have identified is really crucial, or perhaps they

can identify crucial evidence for the prudential reading.[78] In fact, most prudential readings of Hobbes's theory seem to assume that the reply to the fool is a piece of evidence of this sort, although supporters of this interpretation do not use the conceptual apparatus I have introduced in this paper to make their case. However, it remains to be seen whether this passage fulfills the conditions (a)-(d) for legitimately considering it a piece of crucial evidence.[79]

Once we have reached this point, there are further issues to deal with. The most pressing concern revolves around what to do with the evidence for the contrary reading. To complete our interpretation, it seems desirable to explain away those passages one way or the other. Here I will rely on some revisionary work done with Hobbes's reply to the fool, stressing that, insofar this is not the only passage for the prudential interpretation, the following remarks are not enough to explain away Hobbes's different claims that support this account of contractual obligation. Although still a minority, some interpreters argue that we should not take Hobbes's arguments against the fool as introducing his definitive view. The fool does not accept the very idea that moral notions are action-guiding to begin with. For this reason, it is not sensible to overstate Hobbes's prudential reply to the fool; instead, instead, the passage seems best interpreted as providing further reasons to someone who does not care about morals. In other words, the reply to the fool does not say that

[78] Although allowing this point seems to introduce further complications, such as whether there can be multiple pieces of crucial evidence, and if so whether they can point in opposite directions. In fact, if they pointed in opposite directions, it seems reasonable to assume that at least one piece of evidence would fail to be crucial.

[79] For what it is worth, I think the passage does not satisfy conditions (c) and (d), although reasons of space do not allow me to elaborate on this issue. If the revisionary explanation of the fool passage is correct, it also follows that the passage does not satisfy condition (b).

Hobbes really shares the fool's framework of ideas; rather, it introduces an account that an irrational agent can understand.[80]

[80] See e.g. R. Rhodes, "Hobbes's unReasonable Fool," *The Southern Journal of Philosophy*, XXX (1992), 95-97, M. Harvey, "Moral Justification in Hobbes," *Hobbes Studies*, XII (1999), 48, "A Defense of Hobbes's 'Just Man,'" *Hobbes Studies*, XV (2002), 84-85, "Hobbes and the Value of Justice," *The Southern Journal of Philosophy*, XLII (2004), 448, "Teasing a Limited Deontological Theory of Morals out of Hobbes," *The Philosophical Forum*, XXXV (2004), 43, S. A., Lloyd, *Morality in the Philosophy of Thomas Hobbes: Cases in the Law of Nature* (Cambridge, UK: Cambridge University Press, 2009), 296-297 and D. van Mill, *Liberty, Rationality, and Agency in Hobbes's* Leviathan , 132 for different readings along these lines.