ARCHIV FÜR RECHTS- UND SOZIALPHILOSOPHIE, *Online first*
DOI 10.25162/ARSP-2023-0026

ANDERS MOLANDER

Reason and Justice

Hobbes's Dispute with the Fool

ABSTRACT: In *Leviathan*, Thomas Hobbes introduces an imaginary figure, the Fool, who disputes the third law of nature, saying: 'that man perform their covenants made'. According to the Fool, 'there is no such thing as justice'. Also, it is not 'against reason' to break a covenant if it is to one's own advantage to do so. Hobbes claims that the Fool is wrong, but where exactly does the latter's folly lie? Commentators have found Hobbes's answer to be surprisingly vague. This paper examines Hobbes's reply and how commentators have tried to assist him. It argues that Hobbes's vagueness reflects an unresolved tension between – in the words of John Rawls – the 'rational' and the 'reasonable' in his theory, a tension that has in turn led to contradictory interpretations.

Keywords: Hobbes, rationality, state of nature, laws of nature, covenants, social contract, game theory, reciprocity
Schlagworte: Hobbes, Rationalität, Naturzustand, natürliche Rechte, Verträge, Gesellschaftsvertrag, Spieltheorie, Reziprozität

In chapter 15 of *Leviathan* (1651), Thomas Hobbes introduces an imaginary figure whom he calls 'the Foole'. In an echo of Psalms 14:1 and 53:1 – 'The fool hath said in his heart, There is no God' – Hobbes has his Fool declare that 'there is no such thing as justice'.[1] According to him, everyone must look to their own 'self-preservation' and is free to do whatever best serves this end. Breaking a covenant may be wrong; however, so long as it is to one's own advantage, doing so is not contrary to reason.[2]

In his reply to the Fool, the usually clear Hobbes lapses into a vagueness that has perplexed commentators. He says that the Fool is wrong, but where exactly does the latter's folly lie? Is Hobbes's reply consistent with his own premises? Is it the Fool himself who is consistently Hobbesian[3], or does Hobbes in fact have a theory of moral obligation,

1 Thomas Hobbes, *Leviathan: With Selected Variants from the Latin Edition of 1668*. Edited with Introduction and Notes by Edwin Curley, Indianapolis/Cambridge: Hacklett Publishing, 1994, 90. Further references in parenthesis. Curley has modernized spelling and capitalization so that "the Foole" becomes "the fool". However, I have kept the capital letter.

2 On other opponents of justice, see David Gauthier, Three Against Justice: The Foole, the Sensible Knave, and the Lydian Shepherd, in *Moral Dealing: Contract, Ethics, and Reason*, Ithaca & London: Cornell University Press 1990. For Hugo Grotius, the Greek philosopher Carneades plays a similar role to the Fool in Hobbes. See Prolegomena in *De iure belli ac pacis* (1625).

3 See for example Thomas Nagel, Hobbes's Concept of Obligation, *The Philosophical Review*, vol. 68, 1959: 'genuine moral obligation plays no part in *Leviathan* at all, but that what Hobbes calls moral obligation is based exclusively on considerations of rational self-interest' (68); '(Hobbesian man) is susceptible only to

not only one of enlightened self-interest? What makes the passage interesting, aside from the interpretative conundrum itself, is that it involves a fundamental problem in Hobbes's theory of political authority; that is, why do individuals enter into and allow themselves to be bound by a contract that restricts their freedom to pursue their own interests? What prompts individuals who in the state of nature find themselves in a 'war of all against all' to associate, become members of a state and make themselves subject to binding laws?[4]

Let us begin by clarifying the context in which the Fool appears.

I.

In chapter 13, Hobbes describes the 'state of nature' in which human beings exist in the absence of coercive state power. He assumes that they have the same physical and spiritual capabilities, that their primary concern is self-preservation and that they pursue this goal in a rational way, which is to say that they deliberate the means that will most effectively secure its achievement.[5] In the absence of state power, individuals are free to exercise their own discretion about what best promotes self-preservation. At the same time, the fact of scarcity makes conflicts of interest inevitable: 'if any of two men desire the same thing, which nevertheless they cannot both enjoy, they become enemies; and in the way to their end, which is in principle their own conservation, and sometimes their delectation only, endeavour to destroy or subdue one another' (75). The state of nature is therefore a condition of 'war of every man against every man' (76, 78, 80, 84), not in the sense that acts of war are continually occurring but that there is a permanent state of readiness for war. Other people represent an ever-present danger, for which rea-

selfish motivation, and is therefore incapable of any action which could be clearly labeled moral. He might, in fact, be best described as a man without a moral sense'(74). See also Rawls (note 35), 70–71.

4 For example, in *Auch eine Philosophie der Geschichte* Jürgen Habermas states that in Hobbes theory, 'tritt mit besonderer Klarheit das ungelöste Problem des Vernunftrechts hervor …: *die Begründung einer normativen Selbstbindung des Willens freier Bürger*' (Suhrkamp: Berlin, 2019, vol. 2, 144). This way of reading Hobbes with a systematic intent differs from historicising or contextualised interpretations of the kind represented by Quentin Skinner and the Cambridge School (on this methodological program, emphasising the 'performativity of texts', see Skinner, *Visions of Politics. Vol. 1, Regarding Method*, Cambridge: Cambridge University Press, 2002). A prominent example of a 'systematic' reading of *Leviathan* is Jean Hampton's *Hobbes and the Social Contract Tradition* (Cambridge: Cambridge University Press, 1986, 1), which seeks to illuminate 'the general structure of all social contract arguments by analyzing and explaining Hobbes's contractarian argument'. These two interpretative approaches are not mutually exclusive, however, but have different knowledge interests, on the one hand 'rational reconstruction' and on the other to understand Hobbes's arguments as 'speech acts' in a specific intellectual and political context, most immediately that of the English Civil War and the subsequent Cromwellian Protectorate. For a discussion of these various approaches, see Adrian Blau, Methodologies of Interpreting Hobbes, in Sharon A. Lloyd (ed.), *Interpreting Hobbes's Political Philosophy*, Cambridge: Cambridge University Press, 2019.

5 Hampton (note 4), 38, compares Hobbes' concept of reason with what Derek Parfit called 'the Deliberative Theory of Rationality' (see Parfit, *Reasons and Persons*, Oxford: Clarendon Press, 1984, 118): 'What each of us has reason to do is what he would best achieve, not what he *actually* wants, but what he *would* want, at the time of acting, if he had undergone a process of 'ideal deliberation'—if he knew the relevant facts, was thinking clearly, and was free from distorting influences.'

son it is prudent to protect oneself in the best way possible, and there is nothing, Hobbes writes, 'so reasonable as anticipation; that is, by force or wiles to master the persons of all men he can, so long till he see no other power great enough to endanger him' (75).

Industry, trade and culture have no place in this state, but worse still is the 'continual fear and danger of violent death' (76). Human life is, in Hobbes's famous words, 'solitary, poor, nasty, brutish, and short' (76). Nor is it possible for something to be right or wrong, since these terms have no meaning in the absence of a common legislative power. There can thus be no property, 'no *mine* and *thine* distinct' (78), but rather, everyone has a 'right to everything' (80). At the same time, people have passions that predispose them to be peaceful, such as 'fear of death, desire of such things as are necessary to commodious living, and a hope by their industry to obtain them', and, not least, 'reason suggesteth convenient articles of peace upon which men may be drawn to agreement' (78). Hobbes calls these articles 'laws of nature', which, as such, are 'immutable and eternal' (99). A man in the state of nature may seem like 'homo homini lupus' (as a wolf to other people) but has nonetheless social proclivities and, if he can be made to live in accordance with these laws prescribed by reason, can become 'homo homini deus' (as a god to other people).[6] But – and this is Hobbes's core message – it requires 'a common power to keep them all in awe' (76), and thus a change in the conditions under which they interact.

Hobbes describes the freedom universally enjoyed in the absence of state power as 'the right of nature'; that is, 'the liberty each man hath to use his own power, as he will himself, for the preservation of his own nature, that is to say, of his own life, and consequently of doing anything which, in his own judgement and reason, he shall conceive to be the aptest means thereunto' (79). This right does not constitute a claim right; that is, it does not correspond to an obligation (my right to freedom is not something that others are bound to respect, and vice versa). In other words, one's right to freedom does not restrict the same right of others. When the natural right accords everyone the freedom to do what they regard as best to promote their own self-preservation, it creates a state of war that entails misery for all. The 'laws of nature', which set limits to this right to liberty, also have their basis in the interest of self-preservation; a law of nature is 'a precept, or general rule, found out by reason, by which a man is forbidden to do that which is destructive of his life or taketh away the means of preserving the same, and to omit that by which he thinketh it may be best preserved' (79). The same reason that governs individuals in their unfortunate interactions in the state of nature also shows them how this state can be overcome. In both cases, reason instructs them as to what will best serve their self-preservation.

It may seem paradoxical that the natural right to unlimited individual freedom should co-exist with 'immutable and eternal' laws of nature that circumscribe that freedom. Although Hobbes is adopting the terminology of the natural law tradition, he considers it misleading to speak of 'laws'. As 'dictates of reason … they are but conclusions or

6 See Hobbes's dedication, *De Cive*, in Hobbes, *Man and Citizen*, ed. Bernhard Gert, Indianapolis/Cambridge: Hackett Publishing Company, 1991.

theorems concerning what conduceth to the conservation and defence of themselves; whereas law, properly, is the word of him that by right hath command over others' (100). They follow from human beings' natural capacity for reason when applied to the conditions that prevail in the state of nature and have the character of hypothetical imperatives:

> The laws of nature oblige *in foro interno*, that is to say, they bind to a desire they should take place; but *in foro externo*; that is, to the putting them in act, not always. For he that should be modest and tractable, and perform all he promises in such time and place where no man else should do so, should but make himself a prey to others, and procure his own certain ruin, contrary to the ground of all laws of nature, which tend to nature's preservation. (99)

Obligation does not have a moral meaning here. The laws of nature stipulate what everyone should desire, and compliance with them is dependent on what others do. They become actual laws only when they become positivised and can be enforced. Thus far, Hobbes is an early legal positivist, but not in the sense that he excludes an extra-positive source of law, since the legitimising basis of state power consists of the laws of nature that it will make into generally binding positive laws.[7]

The first law of nature enjoins people to strive for and preserve peace; the second directs them to relinquish some of their natural liberty to maintain peace and their own security: 'that a man be willing, when others are so too, as far-forth as for peace and defence of himself he shall think it necessary, to lay down this right to all things; and be contented with so much liberty against other men as he would allow other men against himself' (80). From these two laws follows the third, 'that man perform their covenants made' (89) – that is, the maxim *pacta sunt servanda*. This law is also the basis of the concept of justice as Hobbes understands it: 'whatsoever is not unjust, is *just*.' In the absence of covenants, 'every man has right to everything', and no action can be unjust (89). But even if justice has its origins in the concluding of covenants, it is also the case, according to Hobbes, that 'the names just and unjust … have a place' only when there exists 'some coercive power to compel men equally to the performance of their covenants' (89). Although agreements in the state of nature can be concluded upon the basis of mutual trust, there is always a risk that one of the parties will not uphold them. Stable mutual expectations, and thus valid covenants, only become possible when the third law of nature becomes positive law and is guaranteed by the state's coercive power: 'covenants without the sword are but words, and of no strength to secure a man at all' (106).

In total, Hobbes identifies nineteen laws of nature, of which the fourth to the tenth are various precepts relating to cooperation, and where the tenth prescribes reciprocity: 'no man require to reserve to himself any right which he is not content should be reserved to every one of the rest' (97). The subsequent laws relate to equity and the settlement of dis-

7 See chapter 26, 174: 'The law of nature and the civil law contain each other, and are of equal extent. For the laws of nature, which consist in equity, justice, gratitude, and other moral virtues on these depending, in the condition of mere nature … are not properly laws, but qualities that dispose men to peace and to obedience. When a Commonwealth is once settled, then are they actually laws, and not before, as being then the commands of the Commonwealth, and therefore also civil laws; for it is the sovereign power that obliges men to obey them.'

putes. All the laws, Hobbes writes, can be summarised in 'one easy sum, intelligible even to the meanest capacity; and that is: Do not that to another which thou wouldst not have done to thyself' (99). This maxim is the golden rule, albeit here formulated negatively in reflection of the symmetrical threat situation that prevails in the state of nature.

Although these laws of nature specify the conditions for peaceful co-existence and are insights based on reason, they do not thereby have binding force: 'notwithstanding the laws of nature (which every one hath then kept, when he has the will to keep them, when he can do it safely), if there be no power erected, or not great enough for our security, every man will and may lawfully rely on his own strength and art for caution against all other men' (106). The solution to the general insecurity of the state of nature is a contract by which everyone subordinates themselves to a sovereign who limits their freedom so that it can be made compatible with everyone's desire for security and a 'commodious life'. We have now reached the crux of Hobbes's political philosophy:

> The only way to erect such a common power … is to confer all their power and strength upon one man, or upon one assembly of men, that may reduce all their wills, by plurality of voices, unto one will … This is more than consent, or concord; it is a real unity of them all in one and the same person, made by covenant of every man with every man, in such manner as if every man should say to every man: *I authorise and give up my right of governing myself to this man, or to this assembly of men, on this condition, that thou give up thy right to him, and authorise all his actions in like manner.* This done, the multitude so united in one person is called a Commonwealth; in Latin, *Civitas*. (109)

The individuals who under the conditions of the state of nature cannot be expected to adhere to agreements can nonetheless enter into a covenant that binds them and thus makes possible social cooperation. The state-founding contract is thus also a social contract. At a stroke, it becomes possible for egocentric individuals to cooperate for mutual advantage.

II.

The Fool enters the picture after Hobbes has introduced the third law of nature. He presents him as follows:

> The fool hath said in his heart: 'there is no such thing as justice'; and sometimes also with his tongue, seriously alleging that: 'every man's conservation and contentment being committed to his own care, there could be no reason why every man might not do what he thought conduced thereunto, and therefore also to make or not make, keep or not keep, covenants was not against reason, when it conduced to one's benefit'. (90)

What the Fool denies, Hobbes explains, is not the moral correctness of keeping to agreements but the reasonableness of always following the law of nature: 'he questioneth whether injustice, taking away the fear of God (for the same fool hath said in his heart there is no God), may not sometimes stand with that reason which dictateth to every

man his own good; and particularly then, when it conduceth to such a benefit as shall put a man in a condition to neglect not only the dispraise and revilings, but also the power of other men'(90).[8] According to the Fool, 'all the voluntary actions of men tend to the benefit of themselves; and those actions are most reasonable that conduce most to their ends' (91).

Hobbes finds the Fool's reasoning 'specious' but nevertheless 'false' (91). He begins by restating the claim he made in his outline of the state of nature, namely that there can be no valid covenants: 'For the question is not of promises mutual where there is no security of performance on either side (as when there is no civil power erected over the parties promising), for such promises are no covenants' (91). But the Fool's objection does not relate to the validity of covenants. He argues that it is not contrary to reason to breach the law of nature if it conflicts with one's self-interest – and he invokes an instrumental concept of reason that would seem to be Hobbes's own.

In his reply, Hobbes distinguishes between two instances in which 'it is not against reason' to keep to agreements: 'either where one of the parties has performed already, or where there is a power to make him perform, there is the question whether it be against reason; that is, against the benefit of the other to perform, or not' (91). What has baffled modern commentators is the word 'or'. It may seem unproblematic to claim that there is no conflict with rational self-interest in keeping contracts when there is a state power that sanctions breaches of contracts, since the potential advantages of doing so are outweighed by the costs in the form of punishment, but the Fool would probably then offer a qualification: if the risk of sanctions is small or the potential advantage outweighs the potential risk, why is it not then rational to behave like a free rider?

Although Hobbes begins by setting aside the state of nature – in which there can be no valid covenants – he seems nonetheless to be saying that the Fool's objection is wrong not merely when there is a state power. In the first instance, it is the case that only one of the parties has performed. If A and B have agreed to exchange X for Y, with A transferring X to B, why should B keep his word and transfer Y? What if B saw an advantage to be gained by breaking the agreement? And what would make A be willing to lead the way? A is interested in gaining access to Y but must also be able to expect that B will transfer Y. It would therefore seem as though Hobbes's statement that it is 'not against reason' to keep to agreements when one of the parties has done their bit presupposes that mutually advantageous agreements can generate a mutual trust that makes the agreements binding without the necessity for coercive power. If so, this would contradict Hobbes's description of the state of nature: 'If a covenant be made wherein neither of the parties perform presently, but trust one another, in the condition of mere nature (which is a condition of war of every man against every man) upon any reasonable suspicion, it is void … For he that performeth first has no assurance the other will perform after, because the bonds of words are too weak …' (84). As Jean Hampton observes,

8 Limitations of space here prevent a discussion of the fact that the justice-denying Fool *also* denies God. On the relationship between the 'two fools,' see Michael Byron, Hobbes's Confounding Foole, ed. Lloyd (note 4).

Hobbes seems here to 'adopt the fool's position to explain the failure of contracts'.[9] If the above interpretation of the instance in which one of the parties has performed is correct, namely that 'promises' can become 'covenants' prior to the establishment of coercive power, Hobbes's reply is startling. Does he in fact mean that there is a logic of cooperation and not merely one of conflict in the state of nature?

Hobbes offers two reasons for why keeping covenants is rational. First, if someone does something that is likely to lead to 'his own destruction', that action does not become reasonable if something unexpected should occur that turns the outcome to his own advantage (91). This means that it is not rational to take great risks, even if one should beat the odds and get a personally favourable result. The rationality of an action is determined by what the actor knows or can reasonably expect *ex ante*, not by what happens through pure chance. Second, in the state of nature 'wherein every man to every man … is an enemy, there is no man can hope by his own strength or wit to defend himself from destruction without the help of confederates' (91). All parties expect the same protection from a confederation, and 'he which declares he thinks it reason to deceive those that help him can in reason expect no other means of safety than what can be had from his own single power' (91). The deceiver is dependent for his security on the delusion of others: 'He … that breaketh his covenant, and consequently declareth that he thinks he may with reason do so, cannot be received into any society that unite themselves for peace and defence but by the error of them that receive him' (92). If received, he will remain only for as long as the others do not become aware of 'the danger of their error' (92).[10] But upon such 'errors a man cannot reasonably reckon upon as the means of his security' (92). His downfall is certain 'if he be left, or cast out of society' and he can 'not foresee nor reckon upon' the errors of others (92). From this, it follows that it is 'against the reason of his preservation' not to keep to agreements (92). Hobbes adds that the inclusion of theological considerations would produce the same outcome. However, we will set this aside since he considers the preceeding argument sufficient to show that the Fool is wrong. Nor does Hobbes see any possibility of deriving arguments from 'justice', as what is just coincides with what is rational in the sense that it promotes self-preservation: 'Justice … is a rule of reason by which we are forbidden to do anything destructive to our life, and consequently a law of nature' (92).

Hobbes thus seeks to convince the Fool of what it means to defend the primary interest of one's own security and self-preservation; in the final instance, he has no alternative, since if he is not to perish, he must follow natural law. The argument stands or

9 Hampton (note 4), 65.

10 That the Fool denies justice not only 'in his heart' but 'sometimes also with his tongue' should not be taken to mean that he is openly declaring his intention of breaking agreements, since this would undermine his strategy. Were this so, he would hardly succeed in deceiving others or in being erroneously accepted as a member of their association. By contrast, he is prepared to admit later on 'that he thinks he may with reason do so' (92). Kinch Hoekstra, who claimed that Hobbes only argues against the 'explicit' Fool, observed that the word 'fool' derives from the Latin *follis*, which he translated as 'windbag'. See Hobbes and the Foole, *Political Theory*, vol. 25, 1997. However, the word used in the Latin text of *Leviathan* is not *follis* but *insapiens*, meaning foolish, unwise or ignorant.

falls upon the identification of risks. It might be thought that the Fool accepts that it is not reasonable to do something he knows will be very likely to result in harm to himself, but he does not accept that breaches of contract in general carry the consequences that Hobbes outlines. What he denies is that it is reasonable to make the keeping of agreements into a rule, since there can be instances when it will be to one's own advantage to break a contract. If this is indeed what he is claiming, which is to say that there can be individual cases when reason says something different to the law of nature, the question is whether Hobbes gives a satisfactory reply. That the Fool reserves the right not to keep to agreements in the state of nature should come as no surprise to Hobbes, given his own depiction of the latter. Hobbes nonetheless argues against the Fool by invoking this state in his confederation argument; no-one can survive without confederates, and everyone is dependent for their security upon being accepted into a 'society'. Because he endangers his own existence by alienating himself from others, it makes sense to keep to agreements and likewise to show oneself to be a cooperating partner who is worthy of inclusion in a society. What kind of 'society' Hobbes imagines is unclear, but it seems to be some form of pre-political association for mutual protection. Yet how is such an association possible under the condition in the state of nature of 'diffidence of one another' (75)?

III.

One circumstance that complicates how we understand Hobbes's reply to the Fool is that the opening part of the reply is different in the Latin edition of *Leviathan* (1668).[11] The English translation of the Latin version runs as follows:

> For the question is not of promises mutual in the natural condition of man, where there is no compelling power; for thus those promises would not be covenants. But if there is a compelling power *and* [my emphasis] if one party has performed his promise, the question is then whether the one who deceives does so with reason and in accordance with his good. I say he acts against reason and imprudently. (91)[12]

Crucially, this reformulated passage replaces the earlier disjunctive 'or' with 'and'. Whereas in the 1651 version, it is not contrary to reason to keep one's part of an agree-

11 The 1651 text reads: 'For the question is not of promises mutual where there is no security of performance on either side (as when there is no civil power erected over the parties promising), for such promises are no covenants, but either where one of the parties has performed already, *or* where there is a power to make him perform, there is the question whether it be against reason, that is against the benefit of the other to perform. And I say it is not against reason.' (91; italics mine).

12 The Latin text reads: 'Quaestio enim non est de Promissis mutuis in conditione hominum naturali ubi nulla est Potentia cogens; nam sic Promissa illa pacta non essent; sed existente Potentia, quae cogat, et alter promissum praestiterit, ibi quaestio est, an is, qui fallit cum Ratione, et ad bonum proprium congruenter fallat. Ego vero contra rationem, et imprudenter facere dico.' For the Latin text, see Thomas Hobbes, *Leviathan. The English and Latin Texts*, ed. Noel Malcolm, Oxford: Clarendon Press, 2012, 225. It is also available at https://archive.org/details/leviathansivedemoohobb.

ment if the other party has fulfilled their own obligations *or* when there is a state power that forces them to do so, the later text argues that it is contrary to reason to break agreements if there is a coercive state power *and* the other party has held up their end of the deal. This reformulation exempts the state of nature from Hobbes's disagreement with the Fool such that the previous incompatability with the state of nature disappears. Hobbes is merely telling him what is unreasonable in the state of nature. But is this a sufficient answer to the Fool's objection? As already noted, the self-interested Fool is not opposed to the idea that it is unwise to breach agreements when faced with the threat of sanctions, provided that the individual in question is governed by rational considerations and not a through-and-through 'gambler' or impulsive 'wanton.'[13] But in that case, why does Hobbes introduce him in the first place when his claim would hardly present a challenge to Hobbes's own argument.

Which text should one use: the English edition of 1651 or the Latin of 1668? The latter, claims Pasquale Pasquino, because it removes the passage's ambiguity and makes it consistent with Hobbes's conception of the state of nature as a 'prisoner's dilemma' (something we shall return to shortly).[14] Pasquino also argues that the Latin edition should be treated as authoritative because Hobbes worked on it for many years and because it was with this text that he engaged with the scholarly world:

> The language of the academic community in the 17th century was, indeed, not at all English but Latin, so that if the first version of *Leviathan* was a book published in the context of the civil war for a limited public – the one able to read English – the Latin version is the text that Hobbes worked out in order to address the *communitas doctorum*, also not only an international scholarly public but all of posterity. Thomas Hobbes could not foresee that the marginal language of a Western European island, partially as the consequence and unexpected effect of the religious civil war that took people out of Great Britain, would become a few centuries afterwards the *lingua franca* of the international academic community. Had he been aware of that, he would certainly have spent his time rewriting a second English edition. But Hobbes was a true humanist and he probably believed that Latin would have been forever the language of the *communitas doctorum*.[15]

The publisher of the French edition, François Tricaud, nonetheless claimed that the Latin text was the template for the English one. But this is contradicted by the correspondence that Hobbes, several years after *Levianthan*'s appearance in English, entered into with a young Oxford academic named Henry Stubbe about the latter's translation, which Hobbes later took over, corrected and completed.[16] In a letter of 1667, Hobbes's

13 'Wanton' is Harry Frankfurt's term for a person who lacks preferences of the second category. See Freedom of the Will and the Concept of a Person, *The Journal of Philosophy*, 68, 1971.

14 Pasquale Pasquino, Hobbes, Religion, and Rational Choice: Hobbes's Two Leviathans and the Fool, *Pacific Philosophical Quarterly*, 82, 2001.

15 loc.cit, 408.

16 See Glen Newey, *Hobbes' Leviathan*, Oxon/New York: Routledge, 2014, 43.

Dutch publisher also expressed his delight that Hobbes had completed two-thirds of the translation and was now spending two hours a day on finishing it.[17]

But even if we base our reading on the Latin text, the fact remains that Hobbes, in his answer to the Fool, claims that it is reasonable to be governed by a logic of security and to seek out confederates in a condition in which 'every man to everyman is an enemy'. As he himself notes, confederation presupposes a certain degree of mutual trust: 'every one expects the same defence by the confederation that anyone else does', and therefore 'he which declares he thinks it reason to deceive those that help him can in reason expect no other means of safety than what can be had from his own single power' (91). And without confederates, the Fool is doomed to perish. Sheer self-preservation thus tells him that it is rational to keep to agreements and to show oneself to be trustworthy. To trust that others will mistakenly accept him as a member would be taking too great a risk. Ultimately, then, the Fool has no alternative but to adhere to the law of nature. This argument means that ambiguity as to the character of the state of nature persists; how can it be possible to enter into an association for mutual protection if the state of nature is defined by negative reciprocity or the logic of the prisoner's dilemma?

The same question prompts Hobbes's account of the laws of nature. These are known to people in the state of nature as the dictates of reason, as 'immutable and eternal laws'. They are a means of achieving the peace that 'all men agree … is good' and are thus themselves also 'good' and expressions of 'moral virtues' (100). And yet, as we have seen, Hobbes claims that these laws merely 'oblige *in foro interno*' as a 'desire they should take place'. While this might seem like a curious conception of obligation (not only from a deontological perspective but also in comparison with its ordinary meaning), it accords with Hobbes's description of them as rules of prudence or hypothetical imperatives. They are 'conclusions and theorems concerning what conduceth to the conservation and defence of themselves' (100), and so long as others cannot be relied on to follow them, the Fool's reservation will be rational; the man who would be 'modest and tractable' will 'procure his own certain ruin' (99). On the other hand, it is the case that 'he that having sufficient security that others shall observe the same laws towards him, observes them not himself, seeketh not peace, but war, and consequently the destruction of his nature by violence' (99). The laws of nature prescribed by reason set the conditions for achieving the peace that all desire, but the extent to which they oblige '*in foro externo*' is dependent upon empirical expectations. If there are situations in which one can be 'sufficiently certain' of others' compliance without the existence of state coercion, then the English edition's formulation of the two conditions that must be met for it to be reasonable to keep to agreements would be correct; that is, either one of the parties has already fulfilled their obligation *or* there is a power that forces them to do so. If such is not the case, then the Latin edition's formulation would be the correct one.

17 Thomas Hobbes, *Correspondence*, ed. Noel Malcom, vol. II, Oxford: Clarendon Press, 1994, 693. Quoted in Pasquino, note 14, 416, no. 18.

IV.

To clarify what Hobbes is claiming in his reply to the Fool, we, like several previous commentators,[18] can draw on game theory and its modelling of strategic interaction and rationality; that is to say, of actors who seek to optimise their own benefit in situations in which the consequences of their choices are dependent upon the choices made by other actors with the same goal. There are those who argue that a theory devised three hundred years after *Leviathan* should not be imposed on Hobbes's reasoning,[19] while others contend that the latter is actually an 'exemplary game-theoretical argument'.[20] To be sure, the fact that game theory did not exist in Hobbes's day is not an argument against using it to illuminate the logic of his reasoning.

Let us say that the state of nature has the structure of a generalised (N-persons) prisoner's dilemma (PD),[21] in which actors have to choose between cooperating or not; in our case, it is between acting in accordance with the natural law of performing covenants (and other laws of nature) or deviating from it (them). Everyone benefits from choosing cooperation over non-cooperation, but choosing non-cooperation is more favourable for the individual regardless of what the others do. With two players, it would look like this (the figures denote the ordinal utility of the various outcomes, i. e. how the players rank the different outcomes, with the rankings of Ego first; 1 is best):

		Alter	
		Keep	Deviate
Ego	Keep	2,2	4,1
	Deviate	1,4	3,3

18 See David Gauthier, *The Logic of Leviathan. The Moral and Political Theory of Thomas Hobbes*, Oxford: Oxford University Press, 1979 and *Morals by Agreement*, Oxford: Clarendon Press, 1986, ch. 6; Hampton (note 4); Gregory S. Kavka, *Hobbesian Moral and Political Theory*, Princeton, New Jersey: Princeton University Press, 1986, ch. 4 and The Rationality of Rule-Following: Hobbes's Dispute with The Foole, *Law and Philosophy*, vol. 14, 5–34, 1995.

19 See Newy (note 16), 139–140.

20 Wolfgang Kersting, *Die politische Philosophie des Gesellschaftsvertrag*, Darmstad: Primus Verlag, 1996, 70. Joshia Ober argues that it was ancient Greek philosphers and historians who 'discovered' the theory of rational choice and strategic interaction (see *The Greeks and the Rational: The Discovery of Practical Reason*, Oakland, California: California University Press, 2022).

21 The narrative behind this game is as follows. Two men are suspected of having carried out a burglary together. The police interrogate them separately. There is sufficient evidence to convict them of another, lesser crime, which carries a five-month prison sentence. The prisoners are given an ultimatum. If one of them confesses, he will serve three months in prison while the one who remains silent serves twelve months. If both confess, they each serve ten months. Both prisoners are anxious to serve as little prison time as possible. If both remain silent, they will each get away with five months, but remaining silent will incur a one-year sentence if the other prisoner confesses. Choosing to confess is thus the best option, regardless of what the other prisoner does, incuring either 3 or 10 months in prison versus 5 or 12. Yet both will get a longer sentence than if both had remained silent; that is, 10 months instead of 5.

The best course for Ego is not to keep to agreements when Alter does; the worst is to keep agreements when Alter does not. And vice versa for Alter. The result is a situation in which neither keeps to the agreement. The best symmetric outcome (2,2) is not a stable state, a so-called Nash equilibrium, since both Ego and Alter have an incentive to deviate. However, the worse symmetric outcome (3,3) is stable, because neither of them can choose a better alternative. In other words, the game has a dominant strategy—whatever the other does, "deviate" is the best choice—and Ego and Alter are therefore doomed to end up in the Pareto-dominated state. The Hobbesian solution to this problem of collective action is the establishment of a sovereign who efficiently sanctions breaches of agreement, thereby transforming the interaction structure (i. e. the pay offs). It no longer has the character of a PD. The preferences have been changed and the dominant strategy is now to keep to agreements:

		Alter	
		Keep	Deviate
Ego	Keep	1,1	2,3
	Deviate	3,2	3,3

If the Hobbesian state of nature can be reconstructed as a choice situation of the PD variety, Hobbes ought to concede that the Fool is right; not keeping to agreements is the rational choice. For his part, the Fool, as already noted, should concede that Hobbes is right in that matters are otherwise in a situation in which breaches of covenant are sanctioned by a state power. He should also be able to agree that such a situation, in which compliance with laws of nature in the form of positive laws can be expected, would be collectively advantageous. But would he be willing to bind himself to enter with others into a state-instituting contract and give up some of his freedom to avoid a suboptimal situation? That is, to submit to a sovereign along with everyone else? If the state of nature follows the PD logic, then 'not give up' or 'not submit' must be the dominant strategy, and there can be no way out of the state of nature:

		Alter	
		Give up	Not give up
Ego	Give up	2,2	4,1
	Not give up	1,4	3,3

The Fool's objection would then point to a weakness in Hobbes's contract theory. He makes the state-founding contract appear mysterious; that is, how can rational individuals be able to conclude a contract if the state of nature is a PD situation?[22]

22 'Hobbesian people can keep virtually no contracts, but if so, how can they keep a "social contract" instituting the sovereign? More generally, if Hobbesian people cannot cooperate on much of anything in the state

Yet the contract mystery presumes that it is correct to read Hobbes narratively, as if he is relating the history of exiting the state of nature. Matters are very different if the contract argument is about not the establishing but the legitimising of state power. In that case, the state of nature is meant to serve as a contrast to that of a political condition – merely a thought experiment to show the reasonableness of submitting to a sovereign. For this reason, Hobbes does not mince his words when describing the cruelty of the state of nature. In this reading, the fact that there is no way out of a PD is not his problem. What he is after is a justification for sovereign power. But even if we interpret Hobbes in this way, the question remains whether the mere insight into the mutual benefits of cooperation will motivate individuals to subordinate themselves to the sovereign and allow Leviathan to wield his sword to uphold agreements and ensure that the laws of nature are followed. Whereas the PD logic of the state of nature prevents them from following the dictates of reason, they are enabled to do so by the protection of Leviathan. This means, as Wolfgang Kersting explained, that the contract is simultaneously a social contract and a *Herrschaftsvertrag*. In the same 'logical second', the contract makes individuals both members of society and subjects: 'Der Vertrag ist Grund der Vergesellschaftung der Individuen nur insofern, wie er auch zugleich Grund der Herrschaftseinrichtung ist. Und er besitzt diese herrrschaftsbegründende Funktion nur als eine die Individuen assoziierende und wechselseitig bindende Rechtsfigur. Der vertragliche Zusammenschluss enthält das Modell der Gesellschaft, deren Bestand durch den Leviathan garantiert wird.'[23]

Although the PD interpretation of the state of nature can be reconciled with the view of the latter as a thought experiment and of the contract as a merely a legitimising 'Rechtsfigur', there are several reasons why PD is not an apt formalisation of the state of nature. While the prisoners in a PD cannot communicate with each other and make their choices in isolation, in Hobbes's state of nature, there are mutual promises, albeit not 'valid covenants'. Such promises may just be cheap talk, but the confederation argument that he offers to the Fool requires a degree of mutual trust. What is more, PD is a one-shot game, whereas the state of nature must reasonably include repeated interactions, and what is rational in an one-shot game need not be so in repeated sequences of interactions. One way to interpret Hobbes's reply to the Fool using his confederation argument is that he is claiming the Fool's error lies precisely in reasoning like a participant in a one-shot PD. Had the Fool taken a longer view of his self-preservation and security, instead of focusing upon his immediate personal advantage, he would have seen the advantage of cooperating; that is, to keep agreements when others show themselves willing to cooperate with him.

of nature, how can they cooperate on the sovereign's institution? Unless Hobbes has an effective answer to these questions, his argument collapses, because he will be unable to explain how people escape the state of nature and enter civil society' (Hampton, note 4, 132). Talcott Parsons called this 'the Hobbesian problem of order' in *The Structure of Social Action* (New York: McGraw Hill, 1937) and used it as the starting point for developing his theory of norm-regulated behaviour.

23 Wolfgang Kersting, *Hobbes zur Einführung*, Hamburg: Junius Verlag, 2002, 150

Uncertainty about others' actions is a leitmotif in Hobbes's account of the state of nature. If the problem with the state of nature is that individuals cannot rely upon others being cooperative, not the rationality of defecting per se, the so-called assurance game (AG) would offer a better representation than PD.[24] Unlike the PD game, it has two equilibria: one in which both cooperate and one in which no-one does. Participants have the following preferences over outcomes:

		Alter	
		Keep	Deviate
Ego	Keep	1,1	4,2
	Deviate	2,4	3,3

If cooperating, Ego will get the best result, as is also the case for Alter, but the choice of cooperation rather than deviation presupposes that Ego can trust Alter to cooperate, and vice versa. There is no dominating strategy. The best choice is dependent upon what the actors believe to be the likelihood of others cooperating – or, put differently, what risks they are willing to take. The thorniest case is, of course, when the individual has no clue about the likelihood that the other is willing to cooperate and, as a result, must act under uncertainty and not only risk.

If Hobbes's state of nature can be represented as an AG, his reply is reasonable, since in this situation, keeping promises is not necessarily in conflict with reason, provided that other people – or sufficiently many other people – can be expected to do so. Conceptualising the state of nature as an AG makes it less 'mysterious' that individuals are able to enter into the state-founding contract, provided that we understand contract theory as one that explains how individuals can leave the state of nature. Thus, the task of the sovereign also changes; that is, from changing the PD interaction so that following the laws of nature pays off to giving a public guarantee that everyone can confidently choose the option of compliance that produces mutual advantage.

David Gauthier has suggested that Hobbes should have replied that it is rational to be a 'constrained' rather than a 'straightforward' maximiser'.[25] Whereas the latter maximises expected utility in each individual situation, the former has developed a disposition to choose to cooperate when it is to their mutual advantage and to not exploit potential advantages in defecting.[26] Yet this presupposes that others can be expected to adopt the same disposition and that one can know whether one is interacting with straightforward or constrained maximisers. Gauthier assumes that even if people are not 'transparent',

24 See Michael Moehler, Why Hobbes' State of Nature is Best Modeled by an Assurance Game, *Utilitas*, vol. 21, 2009.

25 See Gauthier, *Morals by Agreement* (note 18), ch. 6.

26 There is not space here to consider Gauthier's own theory of mutual advantage in *Morals by Agreement* (note 18). For Gauthier, a negotiated result is mutually advantageous—and just—if it satisfies 'the principle of minimax relative concession'; that is, if it minimises the relative concessions (as a percentage of the maximum gain) that the participants must make to reach an agreement.

they will at least be 'translucent', so that one has a decent indication of whom one is interacting with. The question is whether the Fool would be satisfied with this reply. He accepts that he may be excluded from beneficial cooperation, but he can nonetheless insist that it is unwise to bind oneself to a cooperative strategy because one can never be certain of how others will behave. Actors may be translucent in small groups, but does this hold for anonymous interactions in large groups? (Hobbes's state of nature can hardly be described as a situation involving interaction in small groups.) For him, the risk of being the one deceived – the 'modest and tractable' party – is too high. And if he can assume that others are disposed to be constrained maximisers, why not be a straightforward maximiser? For the Fool, everything hinges on what empirical consequences follow from a particular course of action.

If Hobbes had replied in the way Gauthier suggests, a further problem would arise in that it would then be unclear why Hobbes needed an absolute sovereign. If it is rational to choose to be moral in general (i. e. a constrained maximiser), Hobbes's politico-legal solution could be replaced by one of 'morals by agreement' based on mutual advantage of the kind that Gauthier advocates in his own contractarian theory. Why submit to a sovereign?

Another attempt to help Hobbes with his answer is offered by George S. Kavka.[27] He focuses on Hobbes's definition of a law of nature as 'a precept, or general rule, found out by reason' (79). Following the laws of nature leads to better outcomes for the individual – in terms of self-preservation and personal well-being – than calculating in each situation what would be most advantageous. According to Kavka, the argument with the Fool might be reconstructed as follows: the Fool claims that 'case-by-case expected benefit reasoning sometimes suggests it would be better, even if unjust, to offensively violate an agreement (or break another law of nature)', to which Hobbes replies that 'the benefits of violation are uncertain, and the risk of failure are so grave' and that it is therefore 'rational, in purely forward-looking terms, to play it safe and follow the generally beneficial third (or other) law of nature'.[28] The Fool's mistake thus lies in the fact that he does not choose the best long-term strategy under conditions of uncertainty, assuming that he wishes to avoid the worst outcome, namely to be ostracised as a person with whom no one is willing to cooperate. It is true that Hobbes's concept of rationality is forward-looking, and, as discussed earlier, the Fool would probably share it. It is also true that Hobbes's logic is minimising; what matters is to avoid harming one's own interests and, ultimately, endangering one's own existence. The key question therefore concerns the degree to which actors make their choices under conditions of uncertainty. The Fool would likely object to Kavka's 'rule-egoistical' version of Hobbes's reply that the degree of uncertainty is generally not such as to make it more rational to bind oneself to the rule than to choose whatever is most advantageous under the circumstances; that is, to adopt a 'case-by-case' approach. Faced with the question of whether it is reasonable for him to be party to a contract that institutes the sovereign who can offer him guarantees about

27 Kavka, *Hobbesian* (note 18), ch. 4 and The Rationality (note 18).
28 Kavka, The Rationality (note 18), 22–23.

the compliance of others, he would ask himself whether he can be sure that a sufficient number of other people are prepared to enter into such a contract.

Even if AG were to offer a better representation of the state of nature than PD, the Fool would still seem to be able to hold his ground and assert himself as a straightforward maximiser by reasoning as if he were dealing with a PD situation. AG, as noted already, has two equilibria, and for a rational actor, even one that looks beyond what is advantageous in any single interaction, the choice of strategy hinges upon what can be expected of others; that is, it is in my interest to adhere to the laws of nature, but only if I can be reasonably sure that others will do likewise.

A third way to assist Hobbes would be to read his reply as an attempt to make the Fool see himself as a participant in repeated interaction. In his famous experiment using repeated PD games, Robert Axelrod showed the most successful strategy to be 'tit for tat';[29] that is, to respond to cooperation with cooperation. If one of the parties has kept their part of the agreement, it is rational for the other party to do the same, as Hobbes says, provided that the latter is dependent on future cooperation. As mentioned earlier, Hobbes's confederation argument can be interpreted as an encouragement to the Fool to cooperate because defection would result in him being excluded from the association; that is, no-one would want to cooperate with him the next round.

The tit-for-tat solution presupposes that the actors in the first round choose cooperation, which requires them to be able to prioritise their long-term interests over a possible immediate personal advantage. Doing so enables a stable cooperation to develop over time. But if cooperation can come about in this way (as Hume later envisaged and as has been developed by game theorists such as Ken Binmore[30]), it is unclear why Hobbes needs the contract that institutes the sovereign. The conflicts within the state of nature must then also be explained by the fact that people do not act rationally but are governed by their passions.[31] For his part, the Fool can stick to his strategy if he imagines a limited number of rounds of the PD game and presumes that the likelihood for a new round is small, something that seems reasonable under the conditions of the state of nature (that is, when life is 'solitary, poor, nasty, brutish, and short'). In that case, it would be rational to break off the cooperation in the final round and so forth all the way back to the first round (so called backward induction). In an infinitely repeated PD game, by contrast, cooperation can elicit cooperation and subsequently establish a Pareto-optimal equilibrium. Yet, as already noted, this would change the character of the state of nature, making it possible to keep to agreements and cooperate without a Leviathan.

29 Robert Axelrod, *The Evolution of Cooperation*, New York: Basic Books, 1984.
30 Ken Binmore, *Natural Justice*, Oxford: Oxford University Press 2005.
31 See Hampton (note 4), 65 ff.

V.

The game-theoretical reconstruction of the dispute with the Fool therefore seems to give him the upper hand – or at least not leave him without a rejoinder. Even with the improvements to Hobbes's reply suggested by his latterday helpers (constrained maximiser or rule-egoism), the Fool could still object that under advantageous circumstances, it may be reasonable to break convenants if it would promote one's own self-preservation. The dispute simply stalls. Introducing justice – for example, that it is wrong to exploit others' good faith – is no trump card because a sense of justice is precisely what the Fool lacks. Hobbes does not do this either. But perhaps he also does not intend to convince the Fool that it is rational to be moral, something that he would understand if he followed the principles of rational choice.[32] Returning to the confederation argument, it is about the Fool making faulty predictions; he believes that he can avoid the disastrous consequences of a broken agreement by systematically deceiving others. But experience, according to Hobbes, shows that this is highly unlikely. To act as if it could pay to break agreements is therefore imprudent, as the Latin version explains: 'anyone who does what, as far as can be foreseen and understood by reason, tends to his own destruction, even though something unforeseen happens which makes the outcome fortunate, has nevertheless acted imprudently, because what happens is unforeseen' (91).

The Fool's foolishness therefore lies in how he regards facts and makes predictions. It is his biased perception of reality that makes him act in a way that runs contrary to the interest of self-preservation. Thus, Hobbes is appealing to his theoretical reason. According to Sharon A. Lloyd, Hobbes's reply can be reconstructed as follows:

1. To be prudent is to form one's expectations by correct extrapolation from past experience.
2. If experience shows that an action can be expected to be harmful, then (even should it turn out well due to unforeseeable events) it is imprudent to expect that the action will be profitable.
3. Experience shows that relying on the errors of others for the success of one's actions can be expected to be harmful.
4. Any expectation that unjust action will be profitable requires reliance on the errors of others.
5. Therefore, it is imprudent to expect that unjust action will be profitable.[33]

If the dispute is viewed in these terms, it does not involve the normativity of the laws of nature. In his reply, Hobbes does not claim anything that would shake these as 'dictates of reason'. He finds the Fool's objection 'specious' because it is based on the premise that 'all the voluntary actions of men tend to the benefit of themselves, and those actions are most reasonable that conduce most to their ends' (91). Yet the Fool combines this premise with a faulty empirical premise and make unreasonable predictions. If he can

32 See Sharon A. Lloyd, Hobbes' Reply to the Foole: A Deflationary Definitional Interpretation, *Hobbes Studies*, Vol. XVIII, 2005 and *Morality in the Philosophy of Thomas Hobbes*, Cambridge: Cambridge University Press, 2009, ch. 4. See also M. Byron, Hobbes's Confounding Foole, in ed. S.A. Lloyd (note 5).

33 Lloyd, Hobbes' (note 32), 56; Lloyd, *Morality* (note 32), 307.

be made to adopt the predictions that Hobbes deems correct, he should realise that it is prudent to keep to agreements, since he would otherwise risk his own destruction. In other words, his fault lies not in thinking that the pursuit of his own advantage is what reason dictates but in believing that the action of breaking agreements can be in accordance with reason. As John Rawls observed: 'Hobbes does not argue contra the fool that the fool appeals to the *wrong* kinds of reasons; he disputes the fool's supposition of *fact*.'[34]

According to Rawls, however, this reply is related not only to the specific argumentative situation – that the Fool is not receptive to arguments about moral obligation – but the fact that Hobbes's concept of practical reason only permits instrumental reasons. In Rawls's distinction between 'the rational' and 'the reasonable', the laws of nature belong to the category of the reasonable, in that they state principles for 'fair social cooperation'. When justifying these laws, however, Hobbes refers only to the rational, or to what lies in the individual's interest, namely self-preservation; the reasonable is justified 'in terms of the Rational.'[35] It may be that Hobbes 'wants to appeal … only to the most fundamental interests which he thinks *none* will question are fundamental' – that he 'drastically simplifies, but intentionally' – but the consequence is that there is no place for 'the *ordinary* notion of moral obligation'; that is, 'a *reasonable* person does not think it is a *sufficient* reason for violating their promise that they thereby gain *some* permanent, long-run advantage.'[36] For Rawls, Hobbes's refusal to refer to the reasonable in his reply therefore attests to the narrowness of his premises; he lacks a concept of moral obligation. Although Rawls noted in passing that the second law of nature expresses a 'principle of reciprocity,' he did not elaborate on this.[37]

VI.

The second law of nature thus reads: if others are willing to give up some of their natural liberty, 'a man [should] be contented with so much liberty against other men as he would allow other men against himself' (80). And, according to Hobbes, all laws of nature can be summarised by the golden rule (99). In light of this rule, they appear reasonable to every individual: 'he has no more to do in learning the laws of nature but, when weighing the actions of other men with his own they seem too heavy, to put them into the other part of the balance, and his own into their place, that his own passions and self-love may add nothing to the weight' (99). These passages show that Hobbes, despite his notion of practical reason, which corresponds to the rational in the sense used by Rawls (and game theorists), invokes a conception of reasonableness.

34 John Rawls, *Lectures on the History of Political Philosophy*, ed. Samuel Freeman, Cambridge Mass.: Harvard University Press, 2007, 70.
35 loc.cit. 55.
36 loc.cit. 70–71.
37 loc.cit. 61.

In her study of Hobbes[38], Sharon Lloyd attempts to revise the standard reading of him as the arch-theorist of instrumental reason. She argues cogently that Hobbes's fundamental normative principle is in fact a principle of reciprocity in the sense of mutual justification. In her view, Hobbes makes the following claim: 'we won't count a man as rational unless he can formulate and is willing to offer, at least *post hoc,* what he regards as justifying reasons for his conduct (and beliefs). But to offer some consideration as justifying commits one to accepting that same consideration as justifying the actions of others, *ceteris paribus*. So one acts against reason when one does what one would judge another unjustified in doing.'[39] Lloyd regards this 'reciprocity theorem', as she calls it, as the fulcrum for Hobbes's derivation of the laws of nature.

Lloyd's interpretation follows a principle of charity.[40] Exerting herself to read Hobbes in such a way as to make his argument as strong as possible, she finds an authentic theory of morality that solves the problem of the normativity of laws of nature. The conventional reading of Hobbes, which focuses on self-interest and instrumental rationality, may have its limitations, but this kind of 'Hobbesianism' ('your father's Hobbes,'[41] as Lloyd puts it) forms an important part of the history of modern political theory. For purely analytical reasons, we cannot avoid it; we would not be able to differentiate between 'contractarianism' and 'contractualism'.[42] A considerable interpretative effort is also required on Lloyd's part in order to rebut 'your father's Hobbes'. While it is hard to disagree with her claim that Hobbes would stand on firmer ground if his position was enriched with reasonableness, to cast him as a precursor to a John Rawls or a Thomas Scanlon is distinctly audacious.[43] It is a matter of textual fact that Hobbes formulates the laws of nature in egocentric terms; they forbid 'a man … to do that which is destructive of his life or taketh away the means of preserving the same, and to omit that by which he thinketh it may be best preserved' (79). Moreover, as we have seen, Hobbes argues that these laws are only binding *foro interno* – as desires – so long as others cannot be expected to obey them. This is reciprocity in its tit-for-tat sense.[44] What is the right thing to do depends upon one's expectations about other people's actual behaviour, in contrast to reciprocity in the sense of a requirement that one should act as one would want everyone else to act.[45] In his formulation of the second law of nature, Hobbes also makes it conditional upon what others are willing to do. Only when a reciprocity of expecta-

38 Lloyd, *Morality* (note 32).
39 Lloyd, *Morality* (note 32) 4; see also 213 and 362.
40 See Donald Davidson, *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, 1984.
41 Lloyd, *Morality* (note 32), 230.
42 That is, contract theories based on 'mutual advantage' (with Hobbesian roots) and 'mutual respect' (with Kantian roots). See Stephen Darwall, Introduction, in *Contractarianism–Contractualism*, ed. Stephen Darwall, Oxford: Blackwell, 2003 or articles in *Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/.
43 See Lloyd, *Morality* (note 32), xiv.
44 Kavka argued that Hobbes had in mind only a 'copper rule' when he quoted the golden rule: 'Do unto others as they do unto you', which 'glitters less brightly as an ideal of moral conduct than does the Golden Rule'. See *Hobbesian* (note 18), 347.
45 On the distinction between these two types of reciprocity, see Alex Worsnip, Hobbes and Normative Egoism, *Archiv für Geschichte der Philosophie* (94) 2015.

tions exists should individuals not demand more freedom for themselves than they allow others to have. The giving up of liberties must be symmetrical for it to be just and generally acceptable. Reciprocity in the sense of mutual justification is thus a criterion that the laws of nature must meet.[46] Lloyd goes further, however, by claiming that they can be derived from the reciprocity theorem. Yet the problem remains that this is not Hobbes's own line of reasoning when he establishes the laws of nature. If Hobbes were to claim that mutual justification is something that we as rational beings demand from each other – something that has the status of a meta-normative principle or superordinate law of nature – it cannot be an expectation-related principle that only obtains *foro externo* when others can also be expected to adhere to it. If so, this would also imply that we have to understand the state of nature as an already moralised state, whereas Hobbes's argument is that a coercive state power is a necessary condition for the laws of nature to be valid.

VII.

Back to the Fool. He would of course not feel obligated to give reciprocal reasons. He is unreasonable and lacks any sense of justice. It is for this reason that Hobbes tries to convince him that it is imprudent not to keep to agreements. Hobbes's human beings, by contrast, are not exclusively rational but also have a sense of justice. They acknowledge the laws of nature *foro interno* and are prepared to limit their freedom to the same extent that others do. Yet the state authority must first create the normative landscape in which it is possible to be reasonable in safety.

To be a citizen is to submit to constraints. But the Fool does not want this. He is therefore a warning figure whom Hobbes invokes to address his readers, namely those people who are already members or citizens of a state. He is not trying to explain how to get out of the state of nature but rather what citizens must accept so as not to end up there. If citizens behave like the Fool, they will undermine the common order that makes it possible to pursue one's own interests in peaceful competition. To be rational, they must also be reasonable. But they cannot achieve this without the sovereign. If this was the lesson that Hobbes wanted to teach, why is he then so vague in his reply to the Fool? The likely answer is that there is an unresolved tension in Hobbes's theory between the rational and the reasonable – a tension that has in turn led to its contradictory interpretation as, on the one hand, game theory *avant la lettre* and, on the other, as a proto-Kantian theory of reciprocal justification.

46 See for example Kersting (note 20).

PROF. ANDERS MOLANDER
Oslo Metropolitan University, Pb. 4 St. Olavs Plass, 0130 Oslo, andersmo@oslomet.no

Acknowledgments: The author would like to thank Nils Gilje, Cathrine Holst, Gaute Torsvik and Anna Engstam for comments and Stephen Donovan for help with the translation of a Swedish draft.