The Productivity of Misreading: Interpreting Hobbes in a Hobbesian Contractarian Perspective

Luc Foisneau

Luc Foisneau, “The Productivity of Misreading : Interpreting Hobbes in an Hobbesian Contractarian Perspective”, in S. A. Lloyd (éd.), *Interpreting Hobbes’s Political Philosophy*, Cambridge, Cambridge University Press, 2019, p. 242-257.

The Productivity of Misreading: Interpreting Hobbes in an Hobbesian Contractarian Perspective

Luc Foisneau

Unlike Kantian contractarian ethics, which puts moral personality at the forefront, Hobbesian contractarian ethics presents itself as the result of an agreement among purely self-interested beings. Introducing a natural obligation to respect the covenants that have been formed among themselves by agents deprived of moral sense has been of utmost interest for Hobbesian contractarians. As David Gauthier explicitly says, “in Hobbes we find the true ancestor of the theory of morality that we shall present[1].” Reading Hobbes as having stripped man of his moral personality, Gauthier wants to demonstrate that even rational egoists need moral norms. Contrary to Kantian contractarian ethics, the Hobbesian theory of morality that he presents is therefore not about the choice of moral principles, but about proving that moral obligation is a condition of rational social cooperation.[2]

There are, of course, different kinds of questions that can be raised by Hobbesian ethical contractarianism. What we shall do in this chapter is consider the type of use of Hobbes’s texts that has been made by the main representative of this school, David Gauthier. We know the importance of game theory in Hobbesian contractarian ethics, but it is also interesting to consider how that sort

[1] Gauthier 1986, 10.

[2] Concerning the moral limits of Hobbesian contractarian ethics, see Hampton in ed. Vallentyne 1991, 46-50.

of ethical theory relies on interpreting particular texts. Although those interpretations can be deemed erroneous, we would like to show how misinterpreting Hobbes has contributed to fruitful inventions in contemporary ethics. More particularly, studying Gauthier's ethical method can help to see how mistaken interpretations of Hobbes have been a source of renewal in contemporary ethics.

In the first part of this chapter, we will show how a misinterpretation of Rawls's *A Theory of Justice* led to the invention of Gauthier's new contractarian ethics, and why that new ethical theory can be described as a Hobbesian one. In the second part, we will show how Gauthier uses Hobbes's answer to the Foole in order to characterize one of the central problems of his own contractarian ethics. The question is: Why should we be obliged to abide by our covenants if it is in our direct interest not to do so? In the last part, we will investigate what the productivity of those various misinterpretations is, both from the perspective of Gauthier's Hobbesian contractarian ethics and, more generally, from the perspective of reading Hobbes.

1. *Between Hobbes and Rawls: Hobbesian contractarian ethics*

In *Morals by Agreement*, David Gauthier uses an interpretation of Hobbes as a resource from which to develop a theory of morality based on a contractual device containing no significant ethical presupposition[3]. In Hobbes's thought, Gauthier finds two central ideas for his contractarian

[3] The reference to Hobbes is present in various chapters of *Morals by Agreements*, but it plays a central part in chapter VI, "Compliance: maximization constrained", where Gauthier offers a developed exposition of Hobbes's argument about justice, famously associated with the discourse of the « Foole » (p. 158-165; we keep the original

ethics: first, that there is nothing naturally 'good' or 'bad' in the actions of an individual[4], and, second, that morality proceeds from an agreement on terms of cooperation. Those two ideas introduce, according to him, a break in a long-standing tradition, that of medieval Christian moral thought, which held that the good and the bad are real qualities, and that morality is about a capacity to act virtuously and not about a capacity to reach agreement with others. The bad reputation of Hobbes's moral theory is due to this transformation, which puts a sense of self-interest in the place of a sense of the good in itself. Though this new ethics reintroduces a moral dimension in the guise of cooperation, its point of departure in self-interest constitutes a real break with the previous natural law tradition: "Hobbes transformed the laws of nature […] into precepts of reason that require each person, *acting in his own interest*, to give up some portion of the liberty with which he seeks his own survival and well-being, provided others do likewise[5]". The latter proviso expresses the contractarian dimension of Hobbes's theory of the law of nature: I accept to renounce some portion of my natural rights on condition that you renounce the same portion of yours. In this picture, morality is not based on autonomous practical reason, nor on a moral sense inherent to the human character, but rather on a contractual device that reduces the chances of conflict among individuals who consider themselves to have a natural right to defend their lives by any means. Further, in Gauthier's theory, the contractarian idea appears as the rational justification of moral constraints when the invisible hand of the market does not function, the perfect market of

---

spelling, « Foole », to indicate the origin of the problem in *Leviathan*). For Gauthier's general interpretation of Hobbes's moral and political philosophy, see Gauthier 1969.

[4] "Perhaps the classic philosophic formulation of a conception of value both subjective and relative was offered in the seventeenth century by Thomas Hobbes." (Gauthier 1986, 51.)

[5] Gauthier 1986, 10. Italics are mine.

classical economists[6] being supposed to operate without any moral constraint on the will to maximize personal utility. Sometimes one has to submit to moral constraints in order to benefit from the advantages that come from social cooperation. One question is about calling those constraints moral[7]: Can those constraints properly be called moral if they are just the expression of our self-interest? Compared to stronger moral imperatives, such as the Kantian one, Gauthier's constraint could indeed be seen as a pale imitation of, if not as a counterfeit, morality.

The questions that must therefore be asked are whether or not we can really speak of morality within the framework of morals by agreement, and, if it is appropriate to do so, what the structure of the moral problem in such a framework is. In particular, what is the point of a Hobbesian contractarian ethics as compared to a Kantian contractarian ethics? In order to answer those questions, one way to proceed is to see how Gauthier found his way into the emerging theories of justice thanks to what may be called a productive misreading of Rawls.

The development of Rawls's theory of justice – from its initial formulation at the end of the 1950s[8] up to the revised version of *A Theory of Justice*[9] – can give us an indication of the difficulty the first critics of this author had in understanding the rupture he caused in the moral tradition. As one

---

[6] "The first conception central to our theory is therefore that of a morally free zone, a context within which the constraints of morality would have no place. The free zone proves to be that habitat familiar to economists, the perfectly competitive market." (Gauthier 1986, 13.) See also Gauthier 1982, 41-54.

[7] "The idea that an initial social bargain would result in a moral principle invites a number of questions. One is why we should take the bargaining outcome to be a *moral* principle." (Gauthier in eds. Gauthier and Sugden 1993, 26.)

[8] "Justice as Fairness", *Philosophical Review*, lxvii (1958), p. 164-194; reprinted in eds. Laslett and Runcinan 1962.

[9] Rawls's subsequent work in *Political Liberalism* to reframe his theory of justice so that it does not rely on the acceptance of a Kantian interpretation is not relevant to the present argument.

of the very first readers of the work, Brian Barry underlined the reasons for their hesitation in his commentary:

Another problem stems from the fact that Rawls seems to have modified his positions during their long gestation period. While initially the emphasis was placed on morality understood as a system of mutual self-defence, analogous, in a sense, with a revised and corrected version of the Hobbesian theory of "natural law", the focus then shifted [i.e. in *A Theory of Justice*] to the desire to act justly being considered as a central part of human development, a natural extension (based on reflection) of love for others and loyalty towards certain particular associations. The desire to be just and to help perpetuate a just society is something of which a man cannot deprive himself without compromising the integrity of his moral being[10].

This, one of the earliest comments on *A Theory of Justice*[11], gives us a good understanding of the tension that existed for the first commentators on Rawls: whether to make an interpretation freely inspired by Hobbes, which premised the just system on rational self-interested agents, or an interpretation inspired by Kant, which conceived Rawlsian contractualism as based on a moral conception of the self. This tension in the interpretation reflects a deeper tension between two understandings of the constraints associated with moral agreements: on the one hand, that which links our moral constraint to follow just rules to our natural duty of justice, as Rawls does in *A Theory of Justice*[12]; on the other hand, that of Gauthier, who aims to deduce moral constraints from

[10] Barry 1973, 2-3.

[11] On Barry's awareness of the distortions he introduced in his interpretation of Rawls, see Barry 1973, 3.

[12] According to Rawls, this link is as follows: "There is nothing inconsistent, or even surprising, in the fact that justice as fairness allows unconditional principles. It suffices to show that the parties in the original position would agree to principles defining the natural duties which as formulated hold unconditionally. We should note that, since

a morally neutral conception of the rational agent. In the former case, man's natural duty is a presupposition, from which it is possible to establish the obligation to obey just institutions[13]; in the latter case, the moral constraints imposed on individuals are supposed to be deduced on the basis of a morally neutral original situation[14].

The transition from making a distinction between these two contractarian ethics in terms of use to a distinction in terms of kind is, without doubt, inevitable, and in some ways enlightening. However, it has the detrimental effect of immediately making Gauthier's theory of mutual advantage and, as a result, the Hobbesian theory that served as its model, appear to be fundamentally lacking in the area of morality[15]. That is precisely the objection that Gauthier wishes to answer with the help of an interpretation of Hobbes's answer to the Foole in Chapter XV of *Leviathan*.

Indeed, Gauthier is well aware that his contractarian theory of morality may also be perceived as a weak theory since in it the morality of agents is not based on natural duties but on a "context of mutual benefit". His point is to show that weakness as a strength, because it reflects a principle of parsimony. Just as Locke said that "a Hobbist … will not easily admit a great many plain duties of

---

the principle of fairness may establish a bond to existing just arrangements, the obligations covered by it can support a tie already present that derives from the natural duty of justice." (Rawls 1971, § 19, 100.)

[13] In *Political Liberalism*, Rawls will still have it that citizens must be stipulated to possess the capacities needed to cooperate on fair terms, of which the ability to have and act from a sense of justice is one. But, clearly, his account of how they come to have that capacity is no longer Kantian.

[14] Moral neutrality is considered by Gauthier as one of the strong features of his Hobbesian contractarianism: "A contractarian theory of morals, developed as part of the theory of rational choice, has evident strengths. […] No alternative account generates morals, as a rational constraint on choice and action, from a non-moral, or morally neutral, base." (Gauthier 1986, 17.)

[15] See Kymlicka 1992, chap. III.

morality[16]", Gauthier observes that the same may be true of the "Hobbist's modern-day successor[17]". That is why, in an article aptly entitled "Between Hobbes and Rawls[18]", he tries to show that it is possible to simultaneously envisage a morally neutral original situation and a moral understanding of the constraints imposed on social interactions through the device of the contract. In support of his argument, and without in any way disputing the strong Kantian dimension of the Rawlsian undertaking[19], Gauthier tries to show his debt to Rawls by recalling the extent to which his theories borrowed from the Hobbesian interpretation of Rawls made in 1977 by Robert P. Wolff in *Understanding Rawls*[20]. Although Gauthier openly acknowledges that Wolff's interpretation "cannot be defended, not in light of Rawls' previous work nor even in light of the totality of *A Theory of Justice*", Gauthier insists that this "does not affect its interest"[21]. On the contrary, we might be tempted to say that this interpretation had the twofold merit in Gauthier's eyes of, on the one hand, coinciding with the way in which he himself had initially understood

[16] Locke MS, quoted in Dunn 1969, 218-19.

[17] Gauthier 1986, 17. For interesting comments on the latter book, see ed. Vallentyne 1991.

[18] Gauthier in eds. Gauthier and Sugden 1993, 24-39.

[19] "We began, or so we thought, with a need for principles of justice to enable persons to escape the generalized prisoner's dilemma of unconstrained interaction. But we find now that according to Rawls's Kantian interpretation, the real need for principles of justice is to enable persons best to express their nature as moral persons in social union with their fellows." (Gauthier in eds. Gauthier and Sugden 1993, 31.)

[20] Wolff 1977.

[21] "It is an interpretation that in the light of Rawls's subsequent work, and indeed, even in the light of the full development of *A Theory of Justice*, cannot be sustained, but that does not affect its interest." (Gauthier in eds. Gauthier and Sugden 1993, 24.)

Rawls' project and, on the other hand, of defining what would become his *Morals by Agreement*[22] project. Referring back to Wolff's misinterpretation of Rawls will therefore allow us to better grasp Gauthier's understanding of the contractarian foundations of his moral theory.

Wolff's initial supposition was, first, that Rawls took as his point of departure the morally neutral idea that men seek happiness and, second, that the notion of a rational device alone was enough to rationally establish a moral philosophy, without appeal to substantive moral convictions. Refusing to have recourse to such convictions, Rawls, according to Wolff, had the idea of building a formal model of society made up of narrowly rational, self-interested, individuals in what contemporary theory of rational choice calls a "bargaining game[23]". Rawls's intuition was that if

he posited a group of individuals whose nature and motives were those usually assumed in contract theory – then with a single additional quasi-formal, substantively empty constraint, he could prove, as a formal theorem in the theory of rational choice, that *the* solution to the bargaining game was a moral principle having the characteristics of constructivity, coherence with our settled moral convictions, and rationality, and making an independent place for the notion of the right while acknowledging the dignity and worth of moral personality.[24]

The essential point is that the result of the Rawlsian bargaining game would be a moral principle by which all the participants would be rationally constrained to abide, *even in cases it would not be*

---

[22] "When I first read Wolff's account, I realized how well he had captured my own initial understanding of Rawls's aim, and how presciently he had characterized what had become my own [i.e., Gauthier's] project in moral theory." (Gauthier in eds. Gauthier and Sugden 1993, 25.)

[23] For one of the origins of the problem, see Nash 1950.

[24] Wolff 1977, 16.

in their self-interest so to do. Wolff's interpretation is very un-Rawlsian, since it ignores the Kantian interpretation of justice as fairness[25], which situates the original position from the start in a moral perspective. However, Wolff's reading of Rawls is surprisingly close to Gauthier's project in *Morals by Agreement*, which can therefore be interpreted, paradoxically, as a Hobbesian version of *A Theory of Justice*[26].

In the context opened up by Wolff's hermeneutic error, morality appears to be the result of three distinct ideas: first, the morally neutral idea of a rational agent; second, the idea of a social contract considered as the result of a rational negotiation among self-interested agents whose motivations, by assumption, are not moral; and, third, the idea that the negotiation results in a moral constraint[27]. In the distorting mirror of this fruitful misinterpretation, the principles governing society, considered here as an undertaking of cooperation with a view to mutual advantage, are the product of a negotiation between rational agents; if the rationality of the agent is measured by her capacity to act without taking into consideration the interests of others, it is easy to see why the premises of Gauthier's argument are said to be morally neutral. The rational agent is a personal utility maximizer and not a moral agent because she is not able to take into consideration the fact that others are not just rational agents following their own interests but also moral persons. The same moral neutrality cannot be attributed, however, to the obligation imposed on the agents to abide by the principles that they have established through a morally neutral negotiation. In fact, while the moral character that *A Theory of Justice* confers on individuals in the original position

[25] *Cf.* Rawls 1971, § 40, 251-257.

[26] Gauthier is explicit on that point : "When I first read Wolff's account, I realized how well he had captured my own initial understanding of Rawls's aim, and how presciently he had characterized what had become my own project in moral theory." (In Gauthier in eds. Gauthier and Sugden 1993, 25.)

[27] Gauthier in eds. Gauthier and Sugden 1993, 25-26.

precludes questioning the status of the obligation to respect the principles of justice – since the resulting choice in favour of the principles of justice reflects an initial moral commitment to abide by those principles –, the situation is different if we deprive Rawls's theory of its Kantian premise, as is the case in Wolff's interpretation. Since it puts to one side the Kantian interpretation of *A Theory of Justice*, *Understanding Rawls* might have been called "Misunderstanding Rawls". But that omission is a breakthrough for Gauthier, since it opens the way to a problem Rawls did not want to raise, and maybe did not see: If we take as our point of departure a morally neutral state of nature, that is, including neither moral persons nor laws of nature, how is it possible that individuals who are only seeking to maximize their advantages should consider themselves *morally* obliged to respect the principles of social cooperation? It is here that Hobbes's moral philosophy takes on its full meaning and begins to resonate with the preoccupations of Gauthier's moral theory.

2. *Interpreting Hobbes' answer to the Foole in a Hobbesian contractarian perspective*

In this second part, it is not my intention to put forward a technical analysis using terms of game theory[28] of Gauthier's answer to the Foole's objection that having agreed to terms of social cooperation does not suffice to create a moral obligation to abide by the agreement. After recalling the key elements of the Foole's objection, I shall show what Gauthier makes of Hobbes's answers

[28] That interpretation is to be found in *Morals by Agreement*, Chap. VI, p. 166-189, starting with the clear statement : "Let us begin our answer to the Foole …" (p. 166).

to that objection and how this interpretation helps both to understand Hobbes and to open the way to a new moral theory.

It is important to emphasize that the objection Hobbes raises is a radical objection[29], which, if he did not refute it, would run the risk of overturning his entire political theory. Granting that justice lies in respect for covenants entered into, and given that the aim of every covenant is the good of the person who agrees to it, might it not sometimes be good for oneself to act unjustly? It is clear that the question concerns neither the beneficial effect of social covenants, which the fool in no way contests, nor the definition of justice as the keeping of covenants, but rather knowing "whether Injustice, taking away the feare of God, (for the same Foole 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 onely the dispraise, and revilings, but also the power of other men[30]". If rationality is the foundation of moral obligation, and if it is sometimes rational not to keep one's covenants, then Hobbes' third law of nature dictating that covenants be kept cannot be a moral obligation, as Hobbes insisted it is.

Gauthier notes a difference of status between the law of nature requiring justice and the preceding laws: the first two laws come from the strong and simple idea that everyone has reason to prefer preservation in a peaceful state to death or wounds in a warlike state. The choice in favor of peace rather than war seems obvious (first law of nature), and since war goes with the refusal to limit one's right to everything, it is also obvious to agree with others to renounce such an unlimited right

[29] The objection begins with an adaptation of a biblical quotation taken from Psalms 14 and 53: "The Foole hath sayd in his heart, there is no such thing as Justice" (Hobbes 2012, 222).

[30] Ibid.

to all things (second law of nature). Both laws are direct answers to the right of nature that expresses, according Gauthier, "a straightforward maximizing view of rational action, subject to the material condition, central to [Hobbes'] psychology, that each seeks above all his own preservation[31]". But the arguments that justify the first two laws do not so obviously apply to the third: "Hobbes recognizes that it is one thing to make an agreement or covenant, quite another to keep it[32]". When Hobbes says that if the third law does not apply we are still in the condition of war, he does not prove that "conformity to it yields any direct benefit"[33]. It is one thing is to say that making an agreement is rational, since "each gains from the mutual renunciation it involves"; but another thing to say that one must abide by it, since "each does not maximize his expected utility in keeping a covenant, in so far as it requires him to refrain from exercising some part of his previous liberty[34]". One may conclude that although it is rational to make an agreement (to get out of the state of war), it is not rational to consider oneself morally obliged to abide by this agreement, if one can benefit by not complying with it. That is precisely the Foole's point when he says "in his heart, there is no such thing as Justice; and sometimes also with his tongue; seriously alleaging, that every mans 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 ones benefit[35]". Gauthier's reformulation of the Foole's objection is true to the original, and the difference from a Rawlsian approach to the question of compliance could not be more stark. Wolff

[31] Gauthier 1986, 159.

[32] Ibid.

[33] Ibid.

[34] Ibid.

[35] Hobbes 2012, 222.

believes rightly that for Rawls, the constraint imposed through the condition of respecting commitments that are undertaken is "so minimal, so natural" that it goes without saying that the obligation that comes with the principles of justice should be respected. In the Kantian interpretation of justice as fairness, the question of obligation is not central because the obligation to act on just principles is part of what having a moral personality, which includes having a sense of justice, means. Hobbes might similarly have considered it obvious that the social contract should be respected, as long as its legitimacy was recognized by each of the contracting parties. However, on Gauthier's interpretation of the Foole's argument, Hobbes rejects any "appeal to moral personality in the *premises* of the contract argument"[36]. Gauthier's interest in the Foole's argument is precisely that he sees it as addressing the "problem of commitment" that Rawls's theory bypassed. If one thinks that men have no moral sense by nature, there is no reason to presuppose that they will automatically respect the commitments they have made, especially if those commitments contradict their immediate interests. It remains to be proved that there really is an obligation to respect agreements, and that is why Gauthier finds it indispensable to turn to interpretation of Hobbes.

Gauthier's takes Hobbes's answer to the Foole to aim to prove our obligation to respect our commitments. If Gauthier puts to one side the first part of Hobbes's reply, which reminds the Foole that the rationality of an action should be measured according to what can be anticipated and not according to an accidental result, it is, I conjecture, because he does not see the part it plays in the second argument[37]. Gauthier interprets Hobbes's second response to the Foole as highlighting the

[36] Gauthier in eds. Gauthier and Sugden 1993, 31.

[37] Gauthier 1971, 161: "Hobbes's first argument reminds the Foole that the rationality of choice depends on expectations, not actual results. It need not detain us." I'll try to show that Gauthier is mistaken in wanting to

fact that participants to the covenant could not be supposed to ignore the Foole's unjust disposition, and so would act on the basis of their knowledge of his deceitful intention. We might say that there is no veil of ignorance as to the intentions of the Foole when he agrees to covenant. Hobbes speaks of a confederacy: The "confederate" is not yet the "subject" of the civil state, but he is part of an association based on the principle of mutual defence, an association that can only fulfil its aim if there is a certain loyalty among its members. The structure of this defence community is not, in fact, that of a republic with a sovereign, but that of a confederation with no central authority – in other words, a pact of non-aggression and mutual defense among individuals who have given up some of their rights in favour of all members but no one person in particular. We can assume that the members of this embryonic community do not entirely escape the logic of the state of nature. The question that remains is whether it is rational to allow into this community a member declaring that "he thinks it reason to deceive those that help him[38]". Hobbes's reply is that such a man "cannot be received into any Society, that unite themselves for Peace and Defence, but by the errour of them that receive him[39]". It is important to specify the nature of the error that would be committed by the confederates were they to admit the unjust man: in the English version of *Leviathan*, it is said that the unjust man intends and sometimes declares to others his intention to violate the law of justice[40]; in the Latin version, one cannot tell whether the Foole is explicit or

---

dissociate the two arguments.

[38] Hobbes 2012, 224.

[39] Ibid.

[40] "The Foole hath sayd in his heart, there is no such thing as Justice; and sometimes also with his tongue" (Hobbes 2012, 222). The reference to the tongue is clearly a condition of publicity.

silent[41]. In the first case – the only one considered by Gauthier, who does not refer to the Latin version – the error is to tolerate within the confederation a person who declares his intention not to respect the agreements he has made; in the second case, the real intentions of the unjust man are not declared, and one can suppose that he keeps his intentions to himself if "in his heart" is implicit in "has sayd (dixit)". This second solution, the Foole keeping silent, would appear more likely: Why would the unjust man shout from the rooftops that he did not intend to keep his word, thereby running the risk of ruining his plans? But the first situation is more interesting for Gauthier, because it allows for a situation of relative transparency of the intentions of the Foole. As this condition of transparency[42] cannot be deduced from the Latin version, it makes the latter version less suitable for Gauthier's argument because his interest is in knowing whether it would be rational to behave as a Foole in a confederacy where the confederates can be aware of the Foole's unjust intentions. The essential Hobbes's passage, quoted by Gauthier, is this:

He therefore 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 errour of them that receive him; nor when he is received, be retained in it, without seeing the

[41] "Dixit Insipiens, Non est Iustitia." (Hobbes 2012, 223.) On the distinction made by Kinch Hoekstra between the silent and the explicit Foole (Hoekstra 1997, 620–654), see Lloyd's discussion in Lloyd 2009, 311-315.

[42] In his game theoretical justification of moral maximization strategies, Gauthier prefers to speak of "translucency": "However, transparency proves to be a stronger assumption than our argument requires. We may appeal instead to a more realistic translucency, supposing that persons are neither transparent nor opaque, so that their disposition to co-operate or not may be ascertained by others, not with certainty, but as more than mere guesswork." (Gauthier 1986, 174.)

danger of their errour; which errours a man cannot reasonably reckon upon as the means of his security.[43]

In order to refute the Foole's second argument, Hobbes has to prove that the Foole acts contrary to that reason which approves (says Gauthier) the direct maximization of self-interest. Why should it be irrational to be a direct maximizer in the situation of confederacy? Because not respecting one's commitment makes the society more fragile in a situation in which such a fragility can be fatal to the confederation. A confederation has no sovereign power to protect it, but depends on its members to protect each other. Respecting the terms of the covenant is, in such a situation, a condition of collective security. The question is whether it is rational to adhere to a direct maximizing conception of rationality within such a confederacy, and Hobbes answers that it is not, because the Foole's success would depend on the irrationality of the other confederates, and it is not rational to assume that everyone else will behave irrationally. To understand this argument, we need to consider a passage that Gauthier does not quote:

[A]nd therefore if he [i.e., the Foole] be left, or cast out of Society, he perisheth; and if he live in Society, it is by the errours of other men, which he could not foresee, nor reckon upon; and consequently against the reason of his preservation; and so all men that contribute not to his destruction, forbear him onely out of ignorance of what is good for themselves.[44]

What makes this argument crucial is that Hobbes considers the argument of the Foole from a social perspective: The irrationality of the Foole's maxim is due to the fact that it does not take others' points of view into consideration, and reasons as if his own actions could only be assessed from a narrowly self-interested perspective. Therefore, the reason for his irrationality – Hobbes speaks of

[43] Hobbes 2012, 224.

[44] Ibid.

an "error" – is that the Foole voluntarily ignores the effects of others' point of view on the outcome of his own action. Not taking others' interests – "what is good for themselves" – into account is not a moral but an epistemic mistake, so to speak, since the Foole acts as if others could not see what he is really up to and could not in consequence retaliate against him for his unjust actions. What Hobbes says is that rationality, in such a situation, requires taking others' interests into consideration, and that not doing so is equivalent to making an "error". But why should it be an error to expect the ignorance of others? After all, it could be said that the Foole is considering that there is some probability that others won't care about his unjust behaviour, or better still, won't even see it. Indeed, as there are many people in a confederacy, not everyone considers others' actions and intentions with equal scrutiny. Since the rationality of the Foole would then be based on probability, one could speak of the Foole's wager: Although he does not hide his unjust intentions, or knows they can be deduced from his actions, he considers that many people won't care, or won't behave according to their knowledge of his being unjust. Hobbes's argument against this position appears in his first response: "[W]hen a man doth a thing, which notwithstanding any thing can be foreseen, and reckoned on, tendeth to his own destruction, howsoever some accident which he could not expect, arriving, may turne it to his benefit; yet such events do not make it reasonably or wisely done.[45]" In the confederacy situation, it is true that the Foole could benefit from the errors of his confederates, but those errors are accidents that could not be "foreseen, and reckoned on", and therefore it is not rational for the Foole to count on them. Even if he succeeds, and some knaves do indeed succeed, his behavior cannot be considered "reasonable" nor "wise", since not taking others' interests into consideration could have been expected to have a negative impact on the Foole's own satisfaction.

[45] Ibid.

The question we now have to tackle is whether Hobbes's second answer to the Foole introduces a change in his conception of rationality. Gauthier thinks it does:

[F]or Hobbes to take full advantage of this response to the Foole, he must revise his conception of rationality, breaking the direct connection between reason and benefit with which he began his reply. Hobbes needs to say that it is rational to perform one's covenant even when performance is not directly to one's benefit, provided that it is to one's benefit to be disposed to perform[46].

One can wonder to what extent such a change – which requires the Foole to be capable of fulfilling his commitments with no direct, immediate benefit – is not simply begging the question. Such a disposition to perform covenants made means that the Foole would be ready to act justly if he can be persuaded that so acting is in his interest. What has to be proved, therefore, is that acting justly is not in contradiction with the interests of the agent but is the best way to contribute to the maximization of his interests. The problem is, that Hobbes does not speak of such a change in rationality, as Gauthier concedes[47], and does not say "that it is rational to perform one's covenant even when performance is not directly to one's benefit, provided that it is to one's benefit to be disposed to perform[48]". Hobbes implies it only indirectly, by refuting the contrary idea that it would be rational to ignore others' interests in having the Foole behave justly. It might be said that the Foole would no longer be a Foole if he could be persuaded that it is rational for him to develop a disposition to justice. His foolishness is linked to his incapacity to see any further than his own immediate interests.

---

[46] Gauthier, 1986, 162.

[47] Gauthier, 1986, 162: "But this he never says. As long as the Foole is allowed to relate reason directly to benefit in performance, rather than to benefit in the disposition to perform, he can escape refutation."

[48] Ibid.

But could the argument for justice be based on a utility-maximizing argument, and if so, would that imply a change in the Hobbesian conception of rationality? Gauthier has it that the conception of rationality identifying reason with immediate benefit is different from the conception of rationality identifying reason with what may be called social benefit – that is, the benefit associated with the disposition to respect covenants so as to be part of a confederacy. But it may be doubted that these conceptions are really different; in both cases rationality is defined in terms of benefits that relate to an individual's interests. There would certainly be a difference in rationality if the social benefit could be considered independently from the individual benefit, but in Gauthier's understanding of Hobbes's second answer to the Foole the social benefit remains associated with the individual's benefit, or to use Hobbes's terms with "the reason of his preservation[49]". The reason for which one should comply with the third law of nature will still be selfish. So, it seems to me, there is no trace of a revision of rationality in Hobbes's second answer to the Foole. The notion of a disposition to justice framed into the process of formation of a peaceful confederation is not Hobbes's but Gauthier's. It is neither to be found in Hobbes nor in Rawls but indebted to the interpretation of both.

What is striking is the fact that Hobbesian rationality goes much further than a simple utility calculus based on the knowledge of one's interests: my actions can only be said to be rational if I consider that others with whom I interact are also capable of making a true analysis of my intentions as far as justice is concerned. The possibility of exiting the Hobbesian state of nature therefore relies on the capacity of interacting agents to act on the knowledge of others' intentions not to be deceitful. A confederacy could thus be seen as the cognitive condition for establishing a civil state, since the confederates are characterized by their capacity to avoid making too many

[49] Ibid.

errors as to the dispositions of others to justice. But that solution is not perceived by Gauthier who looks in another direction, namely, the effect of sovereignty on individual rationality. We shall now examine this other line of argument.

In order to try to bring Hobbes out of what he sees as a deadlock Gauthier suggests that Hobbes "revise[s] his conception of rationality"[50]. Instead of requiring – contrary to the axiological neutrality of the state of nature – a disposition to act morally (for how else to consider the rationality of the disposition to respect one's commitments?), Hobbes conceives, according to Gauthier, a political transformation of right reason resulting from the institution of the State. We should here quote the main text of the polemics between Hobbes and Bramhall, *The Questions Concerning Liberty, Necessity and Chance*, to which Gauthier refers as a key passage:

All the real good, which we call honest and morally virtuous, is that which is not repugnant to the law, civil or natural; for the law is all the right reason we have, and, (though he, as often as it disagreeth with his own reason, deny it), is the infallible rule of moral goodness. The reason whereof is this, that because neither mine nor the Bishop's reason is right reason fit to be a rule of our moral actions, we have therefore set up over ourselves a sovereign governor, and agreed that his laws shall be unto us, whatsoever they be, in the place of right reason[51].

Gauthier deduces from this text that we should answer the Foole by telling him that injustice is not compatible with the reason of the sovereign's law. Just as it is not rational to hold on to one's right to all things when one realizes that it conduces to our destruction, so it is not rational to make use of one's natural reason in the presence of the sovereign when one realizes that the function of the sovereign is to be an artificial right reason. Even so, everyone keeps making use of his natural

[50] Gauthier 1986, 162.
[51] Hobbes 1966, 194.

reason, and at no point does Hobbes say that the sovereign's reason should be internalized by his subjects[52]. What would be the reason for the Foole's obedience – assuming he obeys at all? It is, says Gauthier, the threat of punishment; the solution to the Foole's objection Gauthier puts forward is a political solution, not a moral one. That solution does not solve the problem of the *rationality* of the just action; it only brings into the picture an external motivation to behave justly in the civil state. There is indeed a constraint on the choice of a just action, but that constraint is not an internal constraint linked to the rationality of acting justly, but an external constraint based on the idea that unjust action would result in state punishment.

It turns out that the plan to deduce morality on the basis of axiologically neutral premises is not guaranteed to find in Hobbes such a strong support as Gauthier would have expected. One way of answering Gauthier's interpretation could be to re-evaluate Hobbes's theory of the laws of nature[53]. Indeed, there is no evidence that Hobbes was thinking what Gauthier is supposing him to think – that Hobbes sought to derive the laws of morality from the supposedly axiologically neutral premises of an exclusive concern for self-preservation. Nevertheless, the interest of that misinterpretation lies in the fact that, by attributing to Hobbes an axiological neutrality that is not his, and to Rawls a theory of bargaining he never seriously considered, Gauthier has succeeded in setting out the problem of the origin of moral rules in a new way[54].

---

[52] For a defence of the opposite thesis, see Byron 2015.

[53] Lloyd directs her re-definition of the law of nature against the presupposition that is at the basis of Gauthier's approach to it, "arguing against the common presumption that Hobbes defined a Law of Nature as a precept forbidding an agent to do what is destructive of his own preservation and requiring him to pursue his preservation" (Lloyd 2009, 99).

[54] I want to thank Sharon Lloyd for her suggestions in the process of editing the present chapter.