Economics & Philosophy (2022), **38**, 171–205
doi:10.1017/S0266267121000079

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Unravelling into war: trust and social preferences in Hobbes's state of nature

Alexander Schaefer[1] and Jin-yeong Sohn[2,*]

[1]Philosophy Department, University of Arizona, 213 Social Sciences, 1145 E. South Campus Dr., Tucson, AZ 85721, USA and [2]Institute for Advanced Economic Research, Dongbei University of Finance and Economics, 217 Jianshan St., Dalian, Liaoning, China 116025
*Corresponding author. Email: jinsohn7@dufe.edu.cn

(Received 25 April 2020; revised 2 November 2020; accepted 11 January 2021; first published online 30 July 2021)

Abstract
According to Hobbes, individuals care about their relative standing in a way that shapes their social interactions. To model this aspect of Hobbesian psychology, this paper supposes that agents have *social preferences*, that is, preferences about their comparative resource holdings. Introducing uncertainty regarding the social preferences of others unleashes a process of trust-unravelling, ultimately leading to Hobbes's 'state of war'. This *Trust-unravelling Model* incorporates important features of Hobbes's argument that past models ignore.

Keywords: Hobbes; social contract; contractarianism; game theory

1. Introduction

Hobbes purports to demonstrate that life in the state of nature is 'solitary, poore, nasty, brutish, and short' (Hobbes 1991: 89) – but how, exactly, does he prove this? Beginning with classic works by David Gauthier and Gregory Kavka, several theorists have employed the techniques of game theory to clarify the structure and content of Hobbes's argument. Such an approach seems natural given Hobbes's emphasis on the individual perspective and his insistence that outcomes be deduced from individualistic postulates, just as conclusions in geometry are deduced from definitions and axioms.[1]

This research programme has led to several important insights. It has demonstrated multiple ways in which rational individuals might trap their society in a suboptimal equilibrium, thus showing how we might ground state authority on a purely individualistic foundation. It has added rigour to

[1]Although game theory is especially suitable for analysing Hobbes, political theorists have employed game-theoretical techniques to analyse other social contract theories as well. See, for example, Binmore (2005), Kogelmann and Stich (2016), Kogelmann and Ogden (2018), Thrasher and Vallier (2018) and Chung (2018).

arguments otherwise based on intuition and folk psychology, thus making good on Hobbes's claim to deduce his results by following the methodology of pure geometry. Additionally, and perhaps most importantly, it has raised new questions about the nature of Hobbes's argument that were previously obscured in the vagueness of English prose. For example, does the state of nature pose a coordination problem or a compliance problem? Is preemptive violence a best response because it strictly dominates cooperation or because trust is so elusive?

Despite these advances, new territory remains to be explored. Older models generally exhibit two major shortcomings. First, they often include only one type of player, thus ignoring the diversity that Hobbes explicitly ascribes to individuals in the state of nature. Second, they typically ignore the issue of trust or uncertainty, even though Hobbes identifies 'diffidence of one another' as a key factor driving the emergence of war (Hobbes 1991: 87–88). The model presented here overcomes these simplifications by allowing a full continuum of types and by explicitly incorporating uncertainty as a driver of conflict. Two newer alternative models, those of Vanderschraaf (2006) and Chung (2015), have also succeeded, to some extent, at incorporating uncertainty and diversity. However, as we will see (section 5), our model progresses beyond these models by incorporating their major virtues without exhibiting their major defects.

To present our new model of Hobbes's state of nature, we proceed as follows. In section 2 we survey older game-theoretical treatments, or what we call 'the first wave'. From this critique, five crucial desiderata become apparent. These desiderata, defined in section 3, identify ways to improve upon first wave models in developing an adequate game-theoretical reconstruction of Hobbes's argument. Two other models, those of Vanderschraaf and Chung, have made important advancements in meeting these desiderata. In section 4, therefore, we identify the merits and limitations of these models. With these merits and limitations in mind, section 5 lays out a new model – the *Trust-unravelling Model* – and highlights how it retains the desirable features of the second wave models, while relaxing their limitations. Finally, section 6 contains a summary table of our results, and clarifies an interpretive issue regarding the concept of a 'dominator type'.

2. The first wave

In Hobbes's state of nature, individually rational behaviour produces disastrous results. This conflict between rationality and optimality evokes a game-theoretic analysis. Indeed, formal reconstructions of Hobbes's argument began to appear soon after game theory reached maturity in the late 1950s. While providing a means of elucidating and verifying Hobbes's argument, they also gave rise to disagreements regarding the payoff structure and environmental conditions that Hobbes posits in the state of nature.

According to John Rawls, Hobbes's state of nature furnishes the 'classical example' of the prisoner's dilemma (Rawls 1999: 238). David Gauthier concurs, comparing Hobbes's state of nature to the modern issue of international disarmament (Gauthier 1969: 79–80). According to this view, the sources of conflict – fear, diffidence and glory – entail that the strategy of peaceful

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

cooperation is strictly dominated. In the state of nature, according to this model, each individual strictly prefers to unilaterally aggress. Knowledge of this fact exacerbates the issue, since choosing a cooperative strategy when others choose an aggressive one is utterly calamitous. Facing such an incentive structure, each individual will choose a strategy of preemptive violence. The violent result, viz. *warre*, is universally dispreferred to mutual cooperation.

The one-shot prisoner's dilemma has been criticized for its failure to capture various aspects of Hobbes's depiction of the state of nature. First, the prisoner's dilemma requires that both parties strictly prefer to unilaterally aggress, while Hobbes suggests that his conclusion follows even if individuals in the state of nature 'would be glad to be at ease within modest bounds', preferring universal cooperation over unilateral aggression (Hobbes 1991: 88).[2] This points towards a second issue, i.e. that the one-shot prisoner's dilemma lacks a key feature of Hobbes's state of nature: fear. Even those individuals who prefer universal cooperation are driven to preemptive aggression due to their uncertainty and lack of trust that others are so disposed (or that others, even if so disposed, know that their opponents have similar preferences) (Chung 2015: 488). A third issue, closely connected to the last two, concerns the symmetric nature of a prisoner's dilemma. In both *Leviathan* and *De Cive*, Hobbes asserts that the state of nature may involve agents with varying preferences; for some, aggression is a strictly dominant strategy, while for others, it is simply the safest strategy given that they cannot guarantee that their partner will be peaceful.[3] Hobbes's state of nature is not, or at least need not be, a symmetric game.[4] The failure of the one-shot prisoner's dilemma to account for peaceful preferences, uncertainty, and diversity of types has led to a variety of alternative models.[5]

The iterated prisoner's dilemma stands as a close alternative to its one-shot counterpart. This model, employed extensively by Gregory Kavka (1983, 1986), brings reputational considerations to the fore and provides a more realistic representation of interaction in the state of nature.[6] Despite these virtues, it still exhibits all of the main drawbacks of the one-shot prisoner's dilemma identified above: both parties must prefer unilateral aggression in each isolated interaction, uncertainty plays no role, and players in the game must share homogeneous preferences. Further issues arise when the details of the model must be specified. Is it a finitely repeated prisoner's dilemma? If so, then, assuming common knowledge of rationality, backwards induction implies that mutual defection is the unique Nash

[2]See also Hobbes's discussion in Chapter 11: 'the cause of [power-seeking] is not alwayes ... that he cannot be content with a moderate power: but because he cannot assure the power and means to live well ... without the acquisition of more [power]' (Hobbes 1991: 88)

[3]Importantly, as we will see below (section 3), Hobbes identifies two distinct types in *De Cive*, while in *Leviathan* the types are presented as diverse, but not necessarily a discrete couple. Their preferences may fall along a continuum, a more realistic assumption than that used in *De Cive*.

[4]Strictly speaking, the prisoner's dilemma requires symmetry in preference ranking, but not necessarily in utility payoffs. This point does not affect the argument made here.

[5]Other serious issues with the one-shot prisoner's dilemma have also been noted. For example, if cooperate is never a best response (and hence mutual cooperation is not a Nash equilibrium), then civil society must be unstable (Skyrms 2001).

[6]See also Hampton (1988) and Skyrms (2001).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

Equilibrium, and the finitely repeated prisoner's dilemma becomes virtually indistinguishable from the one-shot prisoner's dilemma (Kavka 1986).[7] On the other hand, if we take the game to be infinitely repeated, as Jean Hampton does, then (assuming reasonably low discount factors) rational players may have no reason to defect and thus have no need for absolute sovereignty. Hampton argues that some players will be short-sighted, not realizing the nature of the game they are playing; they may assume it is a one-shot or a finitely repeated prisoner's dilemma and therefore choose to defect. This, however, implies that violence breaks out only when foolish players are involved in the interaction, a claim that Hobbes seems nowhere to affirm. Instead, Hobbes emphasizes the rationality of anticipatory aggression under conditions of uncertainty: a central reason modest types preemptively attack is because they cannot be sure what type of player they are interacting with (Hobbes 1991: 88). Moreover, although we noted above that rational players *may* have reason to cooperate, it is equally true that they may have reason to defect. With sufficiently low discount rates, the Folk Theorem implies that an infinite variety of strategy profiles might be supported as Nash equilibria. In other words, the infinitely repeated prisoner's dilemma lacks predictive power. It can rule out certain outcomes, but it doesn't predict war; nor does it predict peace. Consequently, it cannot faithfully reconstruct Hobbes's argument, at least as Hobbes sees it, since Hobbes intends to show how war will inevitably arise in a stateless society. There have been some ingenious modifications of the iterated prisoner's dilemma which overcome one or more of these many issues – e.g. by introducing uncertainty into the players' beliefs about whether play will continue or stop in the next period – but none to date have overcome all of these issues.[8]

Michael Moehler has recently provided a related critique of prisoner's dilemma models of Hobbes's state of nature. According to Moehler, since most interactions in the state of nature take place between reasonable individuals (rather than irrational fools), the key problem that must be resolved by the sovereign is not compliance, but assurance. In other words, a rational individual in the state of nature will anticipate, viz. defect, not because she knows that cooperation is a dominated strategy for her opponent, but because she lacks assurance that her cooperative act will be met with cooperation by her opponent. Anticipatory aggression thus provides a safer option. Since the prisoner's dilemma presents defection as always preferred to cooperation, it fails to capture the most important source of violence in Hobbes's state of nature, viz. a lack of trust. In contrast, the assurance game involves a tension between mutual benefit and personal risk; it succeeds in capturing the problem of assurance.[9]

[7]This holds even if players don't know exactly when the game will end, so long as they know there exists some upper bound.

[8]See Skyrms (2001) for an interesting example. The problem of indeterminacy that arises in the infinitely repeated prisoner's dilemma also appears in the next first wave model we discuss: the assurance game. As Skyrms shows, this is no coincidence, since, when infinitely repeated, the prisoner's dilemma becomes, in essence, an assurance game.

[9]In an earlier article, Alexandra (1992) criticizes the prisoner's dilemma and defends the assurance game as a superior alternative for modelling Hobbes's state of nature. However, Moehler's sophisticated textual analysis of Hobbes renders his defence of the assurance game much more compelling than Alexandra's. For this reason, we focus here on Moehler's article.

Moehler's model therefore improves upon the prisoner's dilemma in a significant way. Nevertheless, although it rightly places trust front and centre, his model is too simple to capture important elements of Hobbes's state of nature. First, like the prisoner's dilemma, the assurance dilemma is a symmetric game. It therefore fails to include the diversity of types that Hobbes posits at multiple points (Hobbes 1991: 70, 88; 1998: Ch. 1, section 4). Moehler claims that models involving non-modest types fixate on the issue of compliance, which 'does not arise in Hobbes' state of nature' (Moehler 2009: 313). But he cannot be correct here, since many interactions in the state of nature involve aggressive individuals, who prefer unilateral conquest over bilateral peace.[10] In defence of the assurance model, Moehler suggests that foolish or aggressive types (which he lumps together) are not present in the state of nature proper, but only in what he calls 'the extended state of nature' (Moehler 2009: 313). This distinction, however, goes too far in massaging the text. Aggressive, untrustworthy individuals constitute an important feature of Hobbes's account of the state of nature. They foment fear and thus prompt anticipation (Hobbes 1991: 88). A model that excludes these types is not a model of Hobbes's state of nature. Interestingly, Moehler criticizes another model, the Assurance Dilemma, on the grounds that it models *only* the interactions between a modest type and a defector type.[11] Yet, the same criticism applies to the assurance game, which models *only* interactions between modest types.

A second issue with the assurance game is tethered to its main virtue, viz. its focus on risk and trust. There is only one reason that coordination can break down in an assurance game: neither player knows what move their partner has chosen. Herein lies the essence of a static or simultaneous-move game. Yet, such an assumption fails to accurately depict interactions in a state of nature, where players could communicate, signal their intentions, or move in succession rather than simultaneously – all of which would undermine the likelihood of ending up at the suboptimal equilibrium. The assurance game, therefore, achieves Hobbes's result only by mischaracterizing the nature of human interaction.

This last point may actually be too generous. Despite its unrealistic construal of human interaction, the assurance game still fails to entail Hobbes's conclusion, viz.

[10]Chung refers to these aggressive individuals as 'vainglorious' types, but this term is used by Hobbes in a very specific way. *Vainglory* is a specific type of *glory*, which is a passion defined by Hobbes as a sort of joy arising from thinking about one's own greatness (or 'power and ability'; Hobbes 1991: 42). What makes vainglory special is that one's high self-estimation stems from flattery or conceit, rather than reliable sources, like past experience. Aside from the fact that not all aggressive individuals are likely to be driven by this specific passion, another issue with this moniker is that Hobbes suggests that vainglory will *not* prompt aggression (or else will only prompt aggression towards the weak and defenceless). For, 'Vain-glorious men … are enclined only to ostentation; but not to attempt: Because when danger or difficulty appears, they look for nothing but to have their insufficiency discovered' (Hobbes 1991: 72). In light of these reasons, we avoid the label *vainglorious*, instead calling such individuals either *dominators*, or *unconditional defectors*. As we define them below, these two terms are not synonymous. We identify the formal definitions of these two categories in section 5.2.

[11]For reasons of space, we do not canvass this particular model, as it is less popular than others. For an account of the assurance dilemma, see Kavka (1989) or Moehler (2020).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

warre, since universal cooperation is a Nash equilibrium of the assurance game.[12] This model therefore provides no reason whatsoever to suppose that violent conflict is inevitable.

While recognizing these issues, it is important to keep the virtue of Moehler's approach in sight. The issue with the prisoner's dilemma model, as well as much of the informal commentary on Hobbes, is that it places aggressive dominator types front and centre. Moehler's laudable goal is to shift the focus away from the dominator types and towards the issue of trust. As we will see (section 3), Hobbes's argument for war does not rely heavily on egoism or the presence of dominator types, so Moehler's contribution to the debate offers a helpful new direction. Nevertheless, due to its simplicity, the assurance game exhibits the three issues enumerated above. It is, therefore, lacking as a game-theoretic reconstruction of Hobbes's argument.

To overcome these issues without abandoning Moehler's key insight about trust therefore requires a model that incorporates: (i) multiple types of players, (ii) uncertainty about types and their moves (so that mere communication or sequential moves cannot establish cooperation), and (iii) a unique, violent equilibrium. Along with two recent, sophisticated treatments, discussed in section 4, our model incorporates these elements.

3. Desiderata of the model

The shortcomings of the models constituting the first wave provide a baseline against which newer models must be judged. Any new model should at least be able to improve upon this first group by avoiding their weaknesses while, at the same time, maintaining their strengths. Thus, from this brief survey we can draw out several desiderata that frame the challenge for any new model. As we will see, there are two other models that have made serious headway in meeting these desiderata. Section 5 demonstrates why our model improves upon even these sophisticated attempts to formalize Hobbes's argument.

The first two desiderata emerge from the discussion of the prisoner's dilemma and its iterated counterpart. A major shortcoming of such models is the symmetry of payoffs; both players have a strictly dominant strategy, viz. defect. This does not fit Hobbes's description of the state of nature, wherein many, if not most, 'would be at ease within modest bounds' (Hobbes 1991: 88). A similar problem affects Moehler's assurance game, as well.[13] Ideally, a model should countenance a

[12]The vast literature on equilibrium selection shows that, without communicative or sequential modifications to the game, the choice of equilibrium is totally indeterminate. Harsanyi and Selten (1988) distinguished between 'payoff dominant' equilibria, such as mutual cooperation in the assurance game, and 'risk dominant' equilibria, such as the non-cooperative equilibrium. Experimentalists, e.g. Van Huyck *et al.* (1997), have identified strong path-dependencies in the selection process, while models employing replicator dynamics generally favour the risk dominant equilibrium (Kandori *et al.* 1993). When communication is possible, however, the payoff dominant equilibrium becomes overwhelmingly probable in experimental settings (Cooper *et al.* 1992; Ostrom and Walker 2000: 451–454).

[13]As we will see, the problem even extends, though in a less pointed manner, to Chung's model, where there are only two types.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

spectrum of types. This is both realistic, and more faithful to Hobbes's *Leviathan*, where he makes no clear assumption as to the precise number of types.[14]

Desideratum 1. The model should include a diversity of possible types. This means that, at a minimum, it should distinguish between conditional cooperators and unconditional defectors, and ideally it will include a wide spectrum of possible types, falling along a continuum from altruistic cooperator to egoistic aggressor.

A second shortcoming of prisoner's dilemma models points towards the next desideratum. As argued above, uncertainty is a crucial feature of Hobbes's state of nature, yet the prisoner's dilemma involves no uncertainty whatsoever. No matter what one's opponent chooses, it is rational to aggress and irrational to cooperate.

Desideratum 2. Players should be uncertain as to the type (viz. payoff function) of their opponents, and this should make a significant difference to their choice of strategy.

Hobbes is quite explicit about the role of fear and uncertainty in generating the state of war. As he puts it, 'from ... diffidence of one another, there is no way for any man to secure himself 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' (Hobbes 1991: 87–88). It is *diffidence* that makes anticipation the best response for all individuals.[15] If individuals knew that they confronted a modest type, there would be no fear of exploitation following a cooperative move. If they were certain that they faced an aggressive type, then not diffidence but prudence would favour anticipatory aggression as the best response.[16]

A small clarification regarding the desideratum is needed. We have stated that the model representing the state of nature should ideally have the feature that uncertainty 'makes a significant difference to players' choice of strategy'. To render this more precise, we note that our counterfactual game must be the game where there is no uncertainty. In other words, the desideratum requires that when the game with uncertainty is compared with the identical game *without* uncertainty, players choose different strategies. Later (section 5.4), we will formally define what we mean by 'significant difference'.

The first and second desiderata represent failings of the prisoner's dilemma, iterated prisoner's dilemma, and the assurance game.[17] The next desideratum, by

[14]Some would disagree with this desideratum, citing textual evidence that Hobbes endorsed a two-type model. We engage with this objection in section 4. For the importance of diversity to social contract theory, generally, see Turner and Gaus (2017). For a contemporary application of social contract theory to the problem of diverse private interests, see Delmotte (2020).

[15]*Diffidence* is not clearly defined by Hobbes, but he employs it as a contrast term for both confidence and trust. We follow Alice Ristroph in understanding diffidence as 'uneasiness or anxiety that all individuals ... have about their own security and standing vis-à-vis one another' (Ristroph 2014).

[16]See also *De Cive* (Hobbes 1998: Ch. 1, p. 25): 'Men take precautions because they are afraid.'

[17]Arguably, the assurance game may be thought to satisfy Desideratum 2, in the sense that common knowledge of rationality alone does not provide a unique prediction of the co-player's action. In other words, such uncertainty arises (endogenously) due to the game's dominance *in*solvability. However, Hobbes indicates that some *care* more than others about greater power, and the fact that we don't know the degree to which the opponent cares about power (i.e. the preference of the opponent) is one of the drivers of war. So, while the assurance game involves uncertainty *in actions*, we argue that uncertainty should preferably be modelled as uncertainty in (exogenously determined) preferences. However, to indicate that the assurance game involves a kind of uncertainty that is absent from the prisoner's dilemma, we grant that it *partially* satisfies Desideratum 2.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

contrast, is met by the assurance game. It captures Moehler's important point that the emergence of war does not rely on aggressive types.

Desideratum 3. The model should demonstrate the emergence of war without relying on the presence of *dominator types*.

As worded, this desideratum may leave some room for interpretation. First, we must specify what we mean by 'dominator types'. Dominator types can be informally understood as those types that Hobbes says '[take] pleasure in contemplating their own power in the acts of conquest, which they pursue farther than their security requires' (Hobbes 1991: 88).[18] When we begin to lay out our model (section 5), we will specify a formal interpretation of this type by defining a particular utility function associated with the dominator. Second, for the purposes of formally proving that our model satisfies this desideratum (section 5.4), we adopt the following operational interpretation of Desideratum 3: the model must show that *given Hobbes's background assumptions about human nature, war will emerge as the Nash equilibrium, even in the absence of dominator types*. In effect, this means that even if the proportion of dominator types is arbitrarily low – or even if they are entirely absent – defection remains the best response for *all* types of players.[19]

The motivation for this desideratum is a close textual analysis. Although we have criticized his model on other grounds, Moehler's key insight must be admitted: Hobbes's derivation of the state of war from the state of nature does not rely solely, or even primarily, on the existence of especially aggressive individuals. As evidence for this claim, Moehler points out that Hobbes strictly separates his second law of nature – that all individuals in the state of nature lay down their right of nature and transfer this right to an external authority – from his third law of nature, which exhorts individuals to keep any *bona fide* contracts to which they are a party. For Moehler, this separation indicates two separate problems: (1) the task of leaving the state of nature and (2) the issue of compliance once civil society has been established. The first of these problems concerns assurance: we all want to leave the state of nature, and we're all willing to give up our rights to do so, if only we can be certain that others are likewise disposed. Thus, in the state of nature we do not face a prisoner's dilemma, since the issue is not one of compliance, but of mutual assurance. As Moehler puts it, 'the primary problem of collective action and, strictly speaking, the only problem of collective action that must be solved in Hobbes's state of nature in order for society to be established, is the problem of assurance. The problem of compliance does not arise in Hobbes's state of nature. It arises only after society is established' (Moehler 2009: 309).

Moehler may overstate his case, since he denies the presence of dominator types in the state of nature. To the contrary, Hobbes insists that this type exists in the state of nature and that their presence exacerbates the state of war. Moreover, it seems

[18]This description clearly alludes to Hobbes's definition of *glory*. For reasons discussed above (footnote 2), however, it's doubtful that Hobbes intended to claim that such individuals were moved by the passion of *vainglory*.

[19]When we formalize this desideratum in section 5 it will become clear exactly what we mean by the term 'dominator' and what background assumptions the model requires (section 5.2).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

perfectly plausible that both the second and third laws of nature seek to identify solutions to problems that arise in the state of nature. There may simply be more than one type of collective action problem that occurs in such a state.

Nevertheless, a close reading reveals that dominator types do, in fact, play a surprisingly small role in Hobbes's derivation of the state of war. The true causes of conflict – competition, diffidence, and glory – arise even in modest types. Although many commentators fail to notice this feature of Hobbes's argument, it is patent in Chapters 11 and 13 of *Leviathan*, where Hobbes provides his proof(s) that war arises from the state of nature.[20] In Chapter 11, Hobbes builds upon his analysis of power in Chapter 10, arguing that the zero-sum nature of power, combined with the fact that individuals strive to acquire it in order to secure their long term well-being, entails conflict:

> Competition of Riches, Honour, Command, or other power enclineth to Contention, Enmity, and War: Because the way of one Competitor, to the attaining of his desire, is to kill, subdue, supplant, or repell the other. (Hobbes 1991: 70)

In this chapter, therefore, it seems as if Hobbes has derived a state of war from his basic definitions without needing to introduce the idea of an especially aggressive type. Individuals compete, not because they are aggressive or enjoy dominating others, but simply because, in a zero-sum world, their mere survival requires that they do so.

Similarly, in Chapter 13, where the dominator type first appears, Hobbes arrives at the state of war before discussing the scourge of such types. Hobbes's argument begins by noting the equality of human beings, which generates an equality of 'hope in the attaining of our Ends' (Hobbes 1991: 87). The result is competition for scarce resources, which generates diffidence, or mutual fear. The natural response, according to Hobbes, is to anticipate the acquisitive actions of others. Crucially, this aggressive anticipation, which ensures a state of war, does not implicate any dominator types. Individuals do not anticipate because they prefer unilateral aggression to peaceful cooperation. Instead, anticipatory aggression 'is no more than [one's] own conservation requireth' (Hobbes 1991: 88).

Hence, only after universal anticipation, i.e. a state of active war, has already been shown to arise from facts of human nature and environmental conditions does Hobbes then introduce what we have called the dominator types. Given Hobbes's order of exposition in Chapter 13 – first deriving war and only then introducing the dominator types – such types seem to be an additional reason to expect war to arise. Such types constitute an *exacerbating force*, rather than a *necessary condition*.

[20]Both Gauthier (1969: 17, 21) and Kavka (1986: 97–101) recognize that dominator types play a limited role in Hobbes's derivation. Nevertheless, both contradict their own prose explications in their choice of the prisoner's dilemma as the preferred formal model. Most recent commentators seem to have forgotten altogether what these older commentators recognized: the emergence of war does not depend upon the presence of dominator types, even if they hasten its arrival and exacerbate its effects.

The arguments canvassed above, those appearing in Chapters 11 and 13, draw explicitly upon competition and diffidence. Perhaps the missing cause of quarrel – *glory* – is to be found in the dominator types. This would be a convenient way to understand the role that the dominator types play in Hobbes's argument. It would also be consistent with our interpretation, in which the dominator types are only one source of war, not a necessary element in its derivation. However, *glory*, defined as 'Joy, arising from imagination of a man's own power and ability' (Hobbes 1991: 42) plays a role even in the *modest* cooperator's choice to anticipate:

> every man looketh that his companion should value him, at the same rate he sets upon himselfe: And upon all signes of contempt, or undervaluing, naturally endeavours, as far as he dares (which among them that have no common power to keep them in quiet is far enough to make them destroy each other,) to extort a greater value from his contemners, by dommage; and from others, by the example. (Hobbes 1991: 88)

Thus, glory drives *every man*, not just aggressive types, to pursue reputation as a valuable commodity. One might think that this quotation focuses more on reputation than on glory per se, but Hobbes explicitly links glory to reputation, stating that glory 'maketh men invade ... for Reputation' (Hobbes 1991: 88). Individuals want others to accurately assess their power as measured by their own joyful perception of it, i.e. their *glory*. This should not surprise us. In Hobbes's view, human beings are obsessed with relative status. All individuals resent those who attempt to domineer over them; the dominator types, in particular, however, relish the opportunity to engage in such domineering. Relatedly, all individuals are sensitive to insults. So sensitive that they are willing to risk their lives to defend their reputation in the face of dishonour.[21] This concern for status and reputation is all the more intense without a common power in place to secure peace. As Hobbes alleges to have shown by this point in the text, individuals in the state of nature, whether modest or aggressive, regularly attack one another to maintain access to life-preserving resources and to preempt attacks planned by others. In such a state, those with a reputation for weakness or mercy would present appealing targets, while those with a history of victory and brutality may repel prospective aggressors prior to any actual conflict. To deter aggressors, one must show an ability to defend oneself and to exact revenge. Moreover, a strong reputation may attract allies seeking a protective coalition, thereby contributing to one's power and safety. For these reasons, 'Reputation of Power, is Power; because it draweth with it the adhaerence of those that need protection' (Hobbes 1991: 62). And, as pointed out above, even modest types desire power as a means of preserving their lives.

If competition, diffidence, and glory suffice to generate a state of war, and if these features are present among modest types, as we have argued, then the state of war does not hinge upon the presence of dominator types. Yet, in all versions of the prisoner's dilemma game, the state of war emerges as the (unique) equilibrium only because both players prefer aggression, whether or not their opponent plays cooperatively.

[21]Hobbes was probably correct about this. See Henrich (2017: 117–128, 270–272).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

In contrast to the third desideratum, the fourth and fifth are satisfied by both the one shot and finitely-repeated prisoner's dilemma, but not by the assurance game. As argued above, the assurance game does not ensure that war will arise. To assert that war will arise requires 'extra-deductive' reasoning, that is, reasoning that goes beyond the confines of a priori deduction.[22] There are multiple Nash equilibria in the assurance game, and one of these is universal cooperation. A priori, we cannot, therefore, deduce the emergence of war, as Hobbes alleges to have done. This a priori requirement is important to Hobbes, who professes to follow a geometric method, deductively inferring his conclusions from carefully defined terms (Jesseph 1996: 100). Any argument that cannot deductively show that war arises in a stateless condition cannot, therefore, serve as a faithful reconstruction of Hobbes's argument. This observation gives rise to the final two desiderata:

Desideratum 4. The model should demonstrate that war is the unique equilibrium of the model.

Desideratum 5. The model should derive its equilibrium in a purely deductive, a priori manner, as Hobbes intended.

All five of these desiderata arise from shortcomings in past models, and taken together they provide a touchstone of progress in modelling Hobbes's state of nature. The model developed below (section 5) meets all five desiderata and therefore represents a step forward for game-theoretical Hobbesian scholarship. Before examining this new model, it is instructive to consider two recent models that both offer sophisticated formal reconstructions of the emergence of war in Hobbes's state of nature.

4. The second wave

Having surveyed various attempts to model Hobbes's state of nature and shown how they fail to accurately represent Hobbes's argument, it is now time to consider the two most sophisticated attempts to formally analyse Hobbes's argument: Hun Chung's Bayesian game (Chung 2015) and Peter Vanderschraaf's dynamic simulation (Vanderschraaf 2006). What makes these models superior to those in the first wave is their deft handling of Desideratum 2, which requires the incorporation of uncertainty as a key component driving the emergence of war. In an important sense, the goal of this paper is to carry forward Hun Chung's project of constructing an a priori model of the state of nature while retaining certain desirable features of Vanderschraaf's a posteriori dynamical analysis. In doing so, our model can be favourably compared with both of these alternatives as a reconstruction of Hobbes's argument.

In his static game of incomplete information, Chung provides an elegant a priori analysis of Hobbes's state of nature that emphasizes the role of trust in generating a state of war. We have already seen another model that emphasizes trust, the

[22]In fact, even (non-deductive) computer simulations that allow a population of players to evolve in response to the success of different strategies may fail to select universal defection as the equilibrium. There are elaborate models where universal defection *does* emerge as the unique equilibrium in a population playing the assurance game (Kandori *et al.* 1993), but models with slightly different assumptions show that universal cooperation will also frequently emerge as the equilibrium (Bruner 2015).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

assurance game, but Chung goes far beyond the assurance game by explicitly incorporating uncertainty and endogenous belief formation into his model. The agents in Chung's model are Bayesian belief updaters, uncertain as to whether their opponents are dominators or modest types.[23] Chung shows that either (i) a substantial number of dominator types or (ii) an arbitrarily high value placed on one's life (or security) suffices to ensure a suboptimal state of war as the unique perfect Bayesian equilibrium (Chung 2015: 503).

Despite its many virtues, there are two ways in which Chung's model might be advanced. First, Chung assumes the existence of only two types. This idealization runs afoul of Desideratum 1, which calls for a wide spectrum of possible types. Although the incorporation of two distinct types places Chung's model ahead of the prisoner's dilemma and the assurance game, since these models include only one type, a mere two types is neither realistic nor faithful to Hobbes's argument in *Leviathan*. Chung provides textual support for this assumption by citing two different passages. One comes from *De Cive* and does, indeed, assert that we can divide human beings into those who are 'modest' and those who are 'vainglorious' (Hobbes 1998: Ch. 1, section 4). The other is the familiar passage from *Leviathan* where Hobbes states that some are content to stay within 'modest bounds', while other find intrinsic pleasure in 'acts of conquest' (Hobbes 1991: 88).

Evidence from *De Cive* must be used cautiously. It cannot, on its own, justify the strong assumption that there are only two types. This is because the argument in *De Cive* differs in many ways from that in *Leviathan*, including, even, its basic order of exposition.[24] In addition, Kavka suggests that Hobbes may have changed his views on both human nature and the prevalence of dominator types between the writing of *De Cive* and *Leviathan* (Kavka 1986: 99, fn. 37).

Given that *Leviathan* is Hobbes's mature, considered political doctrine, then, it's important that the other passage is drawn from this work, rather than *De Cive*. However, the second passage does not distinguish strictly between two types; it simply asserts that some will seek power 'further than their security requires' (Hobbes 1991: 88). But this is compatible with an infinite variety of payoff functions, not a mere two. Or as Vanderschraaf puts it, 'there are no good reasons to suppose that the moderates or the dominators in anarchy share a single payoff function over alternative outcomes [just] because they share a preference set over these outcomes' (Vanderschraaf 2006: 258). A similar ranking, in other words, may underlie an infinite variety of utility functions.

The second way to advance Chung's project concerns Desideratum 3, namely that the emergence of war does not rely on the presence of dominator types. Chung acknowledges the danger of relying too heavily on dominator types (Chung 2015: 490), while also recognizing the importance of showing that a state of war is the unique equilibrium. Indeed, the key premise that Chung derives in order to prove that war is inevitable is his Lemma: 'In the state of

[23]Chung uses the term 'vainglorious type' in lieu of 'dominator'. For reasons stressed above (footnote 2), we prefer the terms 'dominator' or 'unconditional defector'.

[24]For a detailed discussion of the salient differences between these two arguments, see Gauthier (1969: 34–35).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

nature, launching a preemptive attack is the dominant strategy for everybody regardless of his/her type' (Chung 2015: 489). Chung demonstrates that the Lemma is satisfied in two possible ways: either (i) there is a significant proportion of dominator types in the population (Chung's proposition 2) (Chung 2015: 500),[25] or (ii) individuals in the state of nature value their lives to an arbitrarily high degree (Chung's proposition 4) (Chung 2015: 502).[26] Option (i) clearly runs afoul of Desideratum 3: as we have seen, an accurate model of Hobbes's state of nature should produce a state of war without the presence of dominator types. And as we will see, Peter Vanderschraaf's result requires only an insignificant proportion of dominator types. Again, the importance of this Desideratum lies in Hobbes's own presentation of his argument, as detailed in section 3.

If option (i) proves unsatisfactory, can Chung seek recourse in option (ii)? Although Chung offers a defence of this assumption, there are two decisive objections to assuming an arbitrarily high valuation of life. The first is textual. To defend his assumption, Chung presents a quotation from *De Homine* where Hobbes states that 'the greatest of goods for each is his own preservation'.[27] Aside from the risks associated with drawing quotations from outside works in order to characterize Hobbes's argument in *Leviathan*, this passage also fails to establish that survival is *lexically prior* to all other goods. As Kavka notes, a more plausible interpretation is that 'death is worse than any other *single* evil ... but will not exceed all *combinations* of other evils'.[28] Indeed, in the very same work that Chung quotes, Hobbes writes that

> Though death is the greatest of all evils ... the pains of life can be so great that, unless their quick end is foreseen, they may lead men to number death among the goods. (Hobbes 1998: Ch. 11, section 6, 48–49)[29]

So, the valuation placed on life, even if high, is not arbitrarily high. There is some finite value that agents place on the preservation of their lives, and this necessitates that a non-arbitrary proportion of dominator types be present in order for Chung's Lemma to hold true.

The second major problem with the assumption of an arbitrarily high life valuation is that it conflicts with one of Chung's major goals, namely to establish the inevitability of war without relying on psychological egoism. Chung argues that philosophers have wrongly attributed psychological egoism to Hobbes, and, consequently, misinterpreted his argument for the emergence of war. 'Our model', Chung writes, 'has the advantage of explaining the universal conflict in the state of nature without assuming that everybody has a strictly egoistic psychology' (Chung 2015: 506).[30] Saying that individuals value their lives very highly is, of

[25]See also Chung's supplementary appendix (2015: 9–10).
[26]See also Chung's supplementary appendix (2015: 12–13).
[27]Hobbes (1998: Ch. 11, section 6), quoted in Chung (2015: 504).
[28]See Kavka (1986: 81).
[29]See also Kavka (1986: 81).
[30]Elsewhere, Chung has forcefully defended the claim that Hobbes was not an egoist (Chung 2016).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

course, compatible with rejecting psychological egoism.[31] However, saying that individuals value their own lives at an *arbitrarily* high value – the claim needed in order to prove the Lemma with an arbitrarily low proportion of dominator types – is not compatible with rejecting egoism. An arbitrarily high life valuation means that individuals will reject any choice, including any altruistic action, that places even a slight risk on their own lives. This is, of course, highly implausible as a psychological claim. But more importantly, it undermines Chung's defence of a non-egoistic reconstruction of Hobbes's argument. This second route towards proving the Lemma must therefore be rejected.

To summarize, in order to prove the crucial Lemma and thereby satisfy Desideratum 4, Chung must either assume the presence of a significant proportion of dominator types (violating Desideratum 3) or he must accept a premise that entails the egoistic interpretation that he so vehemently rejects. Chung recognizes the issues with both of these assumptions. Given the intensity with which he rejects psychological egoism, however, he would likely prefer option (i) to option (ii). But this commits Chung to assuming a significant proportion of dominator types, in opposition to Hobbes's presentation of his own argument. Chung's model, while enlightening in several ways, thus fails to satisfy Desideratum 3.

Nevertheless, in contrast to the prisoner's dilemma, Chung's model relies on a small proportion of dominator types. This is due to the fact that individuals in the state of nature value life very highly (though not infinitely). Consequently, according to Chung's analysis, a small number of dominator types suffices to spark war. Although Chung ultimately requires a non-trivial number of dominator types to deduce the emergence of war, the fact that he relies on a small proportion represents a significant difference between his model and the prisoner's dilemma. To mark this relative superiority, we consider Chung to have *partially* satisfied Desideratum 3.

The other model of the second wave, Vanderschraaf's 'variable anticipation threshold model', stands in stark contrast to Chung's static, Bayesian game. Vanderschraaf develops a dynamic, evolutionary model of the state of nature. While dropping the assumptions of common knowledge and homogeneity of types, Vanderschraaf shows that an arbitrarily small number of dominator types suffice to generate a condition of war. Players are assigned preferences that vary randomly. They then begin interacting blindly with other players, updating their strategy as they learn from experience what provides the highest payoffs. Whenever there are dominator types in the population, even just a tiny proportion, the tendency is towards *anticipation*: that is, towards defection, even for the modest types (who rank mutual cooperation above unilateral defection).

Notice how Vanderschraaf's model avoids both of the limitations that Chung's model faces. First, although Vanderschraaf classes players into two categories, Vanderschraaf's model does not posit that there are only two types in the state of nature. Indeed, as Vanderschraaf points out, the fact that moderate types and dominator types each share a preference *ranking* over the outcomes does not imply that they must share a single payoff function (Vanderschraaf 2006: 258).

[31]This is, in fact, an aspect of Kavka's predominant egoism (Kavka 1986: 64–80).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

It is both more general and more realistic to allow a spectrum of payoff functions, rather than stipulating a mere two. It is also more faithful to the text of *Leviathan*, where Hobbes commits himself only to the claim that some prefer unilateral aggression to mutual cooperation. Second, Vanderschraaf also shows that an arbitrarily small proportion of dominators suffices to spark a process of degeneration leading to universal war. This occurs without positing that players hold an arbitrarily high life valuation.

Vanderschraaf's model therefore satisfies Desideratum 1 by including a wide array of possible player types (i.e. of possible utility functions). His model also involves uncertainty, which is captured by the fact that he makes no assumption of common knowledge, either of rationality or of players' types. This uncertainty plays an important role, since it drives agents to employ a myopic learning rule to determine their next move. Desideratum 2 is thus well handled by Vanderschraaf's model. The model also avoids *heavy* reliance on dominator types. It does require such types, but his result follows even if they constitute an infinitesimal proportion of the total population. So, although desideratum 3 is not fully satisfied, Vanderschraaf's model comes closer than Chung's to satisfying Desideratum 3. Moreover, Vanderschraaf succeeds in revealing that war is the overwhelmingly likely outcome in the state of nature, so long as some – even a miniscule proportion – of individuals are dominator types.

Despite these virtues, Vanderschraaf's model misses certain aspects of Hobbes's argument. The first issue, mentioned in the preceding paragraph, is that it involves a retreat from Moehler's insight that even prudent, modest individuals – those prizing universal cooperation above unilateral aggression – will fail to achieve a socially optimal outcome. In Hobbes's argument, the dominator types serve to exacerbate the problem of achieving cooperation, but are not the fundamental cause of war. The fundamental causes – fear, diffidence and glory – arise and persist even in the absence of dominator types. As argued above (section 3), Hobbes makes this quite clear. In contrast, Vanderschraaf's model posits the presence of dominators as 'both a sufficient and a necessary condition to destabilize the [cooperative] state and drive the system to the ... equilibrium of war' (Vanderschraaf 2006: 269).[32] In other words: no dominators, no war. This limitation means that Vanderschraaf's model does not fully satisfy Desideratum 3.[33]

A second limitation concerns the modelling technique that Vanderschraaf employs. We noted above that Vanderschraaf's model reveals the state of war to be the overwhelmingly likely outcome. This does not, however, satisfy Desideratum 5, because Vanderschraaf does not demonstrate this result deductively, as Hobbes would have it. Instead, he employs a dynamical simulation. Such computer-based simulations are empirical in nature. They set up the game by determining attributes of players (their payoff functions, strategy sets, initial strategies, and update rules) and the network determining the patterns of interaction. The theorist then observes

[32]This biconditional holds 'for a wide range of model parameters', but not for all parameters. For instance, if modest types accidentally behave aggressively with a high probability, then dominators are not required for war to arise.

[33]It is worth noting, however, that with respect to Desideratum 3, both Vanderschraaf's model and Chung's *greatly* outperform the prisoner's dilemma models of the first wave.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

what outcome emerges, rather than deducing the equilibrium outcome from an initial specification of the game. This approach contrasts greatly with that of Hobbes, who sought to provide something like a *geometry of civil philosophy*, in which we 'begin [our] ratiocination from the Definitions, or Explications of the names we are to use' and, beginning at these definitions, 'proceed from one consequence to another' (Hobbes 1991: 33–34). In Hobbes's view, true science proceeds by carefully laying out key definitions and deducing conclusions therefrom. In constructing a civil philosophy, he seeks to follow this exacting method.[34] Thus, although Vanderschraaf arrives at Hobbes's conclusion (ignoring the qualification introduced above), he does so in a way that does not track Hobbes's own argument, or even his basic method. Vanderschraaf actually affirms this, stating that his model embodies a dynamical, 'Humean approach' that he takes as superior to an a priori approach like that of Hobbes (Vanderschraaf 2006: 271). Given Vanderschraaf's recognition of this, it is hardly a criticism of his model *as a model*. It is, rather, a criticism of his model *as a reconstruction of Hobbes's argument*.

The third limitation of Vanderschraaf's model also concerns his modelling technique. In general, game theorists prefer an a priori approach to computer simulations for various reasons. First, because a priori analysis proceeds via deductive steps, its result is obtained with absolute certainty, rather than being observed as a mere statistical regularity. Second, the rationality assumption that a priori analysis employs allows explicit proofs to be laid out, leading to full transparency and, sometimes, greater intuition. Finally, a priori proofs avoid the thorny question about the initial population state. The results of simulations can vary dramatically based on the initial configuration of population traits, e.g. the proportion of dominator types. Without a strong justification for one configuration over another, these simulations will lack clear implications.

It should be noted, however, that Vanderschraaf's model escapes the last of these issues, since he observes that *any* proportion of dominator types suffices to generate war. Nevertheless, he does face the first two drawbacks. That said, Vanderschraaf rightly points out that the benefits of certainty and transparency exhibited by a priori analysis do not come for free. This kind of analysis requires the assumption of common rationality, which Vanderschraaf considers to be unrealistic in the state of nature. When we have compelling reasons for relaxing this strong assumption, then dynamical simulations based on simple decision heuristics provide a valuable alternative. However, if the goal is to reconstruct Hobbes's own argument, we actually have strong reason to hold fast to the a priori approach, rather than Vanderschraaf's simulation-based approach.[35]

In sum, while both second-wave models provide a clear and rigorous route to the state of war, and, even more importantly, both models incorporate uncertainty as a driving force towards universal defection, there remains room for advancement. More specifically, Chung and Vanderschraaf's models present the following

[34]For a deeper discussion of Hobbes's methodology, his view of science, and its relation to civil philosophy, see Jesseph (1996: especially 86–87).

[35]Moreover, the extremely simplistic and myopic learning heuristics employed in simulations like Vanderschraaf's strike us as quite unrealistic. Perhaps as unrealistic as perfect rationality, at least in some cases.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

challenge: can a standard game-theoretic model, rather than a dynamic simulation, accurately represent Hobbes's argument without relying on a significant proportion of dominator types in the population? Without losing sight of the fundamental importance of trust – Moehler's insight – the *Trust-unravelling Model* meets this challenge. Instead of relying on a set number of types with defined payoff functions, the Trust-unravelling Model derives its results from a set of weak assumptions, leaving it open to an extremely wide variety of specifications with respect to the beliefs and preferences of the players. This allows the a priori derivation of Hobbes's result without relying on computer simulations, like Vanderschraaf does, and without relying on a significant proportion of dominator types, as Chung does. In what follows, this new model takes centre stage.

5. The Trust-unravelling Model

5.1. Prisoner's dilemma revisited

As argued above, the prisoner's dilemma *game* cannot successfully reconstruct Hobbes's argument a priori. So, the literature has considered completely different games as alternatives. However, by carefully studying the prisoner's dilemma *structure*, we find that it can, in fact, aid us in constructing a logically consistent and textually faithful formalization of Hobbes's state of nature.

Consider one version of the prisoner's dilemma game in Figure 1.

Before we dive into the analysis, one clarification is essential. That is, what do the numbers in the bi-matrix signify? Often in the literature on formal political theories, the numbers, also referred to as the payoffs, are taken to model two distinct aspects of the game. First and foremost, they represent players' decision tendencies, that is, their preferences.[36] And players are assumed to choose an action that gives the highest payoff. Due to this interpretation of the payoff, scholars take for granted that the players must choose the action, D; D results in a larger payoff, regardless of the opponent's action.

Second, the numbers are also taken to represent the material payoff (or 'resources', for short) of the players. This explains why scholars have associated mutual defection with the state of war, and mutual cooperation with peace: $(2, 2)$ is Pareto-superior to $(1, 1)$ (Chung 2015: 490).

Note that the players' decisions need not maximize the players' resources. In fact, Hobbes himself seems to implicitly assume that players do not always strive to maximize their own resources, when discussing the modest types.[37] Some individuals are 'glad to be at ease within modest bounds' (Hobbes 1991: 88). We interpret Hobbes to mean that such individuals would prefer C, when the co-player chooses C, despite the potential *material* gain from D. This means that, even if maximizing their preferences, they are not maximizing their material payoff. In this case, the modest type's decision preference is not represented in the matrix.

[36]Strictly speaking, preference and choice must be considered as distinct concepts. However, for the purposes of this paper, we loosely equate the two. In other contexts, this may give rise to a host of confusions. For a discussion of this issue, see Hausman (1992: 19–22, 2011: Ch. 3) or Lehtinen (2011).

[37]Some commentators have taken this as evidence that Hobbes was not a psychological egoist. See, for example, Chung (2016) or Barrett (2020).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

		Player 2	
		C	D
Player 1	Cooperate (C)	2, 2	0, 3
	Defect (D)	3, 0	1, 1

Figure 1. A version of prisoner's dilemma.

The key insight to be highlighted here is that most models have assumed without justification that the two aspects of the game coincide, but we will show that by separating these two aspects, we are able to construct a model that better represents the preferences of individuals in Hobbes's state of nature. This is precisely what we do in the Trust-unravelling Model. Distinguishing between preferences and material payoffs, we will show, allows us to model the state of nature in a way that satisfies *all* of the desiderata.

Now, consider the following situation: two players, player 1 and player 2 simultaneously choose to either cooperate (C) or defect (D). If both choose C, they each earn $2. If one cooperates and the other defects, the cooperator earns $0 and the defector earns $3. If both defect, they earn $1 each. This structure is summarized in Figure 2.

Following the literature, we interpret the sub-optimal outcome, ($1, $1) as the state of war, and the Pareto optimal outcome, ($2, $2) as the peaceful state.[38] Note that the dollar representation is simply for the ease of exposition. This can be replaced by any limited resources that citizens may fight over: wealth, honour, safety, etc. – all of which Hobbes lumps under the heading of 'power' (Hobbes 1991: 62–64).

Clearly, this structure resembles the prisoner's dilemma *game*. But, in fact, this is not yet a game, for we have not specified the *preferences* of each player. Such a structure, one with these material payoffs but unspecified preferences, is called the prisoner's dilemma *game form* (PDGF). This game form would indeed become the prisoner's dilemma *game*, in the traditional sense, if the players are assumed to be perfectly selfish (in terms of the material payoff).[39]

For the sake of modelling the state of nature, assuming purely self-interested preferences is unnatural for several reasons, many of which we have already pointed out. First, such an assumption would fail to represent the modest types. Second, assuming common knowledge of player types, the assumption of universal egoism would eliminate all uncertainty or diffidence, since all players would defect all the time. Third, Hobbes's view on this matter is ambiguous.[40] For present purposes, we remain agnostic as to Hobbes's psychological assumptions. As we will see (section 5.5), however, the Trust-unravelling Model does place certain psychological restrictions on populations in order to derive

[38](C,C) is not *uniquely* Pareto optimal, since (D,C) and (C,D) are technically Pareto optimal, as well.

[39]It is important not to overthink this standard distinction. A *game* represents the all-things-considered utility payoffs of given outcomes, while a *game form* represents some of the payouts that serve as inputs to players' utility functions. Utility need not increase monotonically in these inputs. This means that a prisoner's dilemma *game form* need not generate a prisoner's dilemma *game*. We thank an anonymous reviewer for pressing us to clarify this distinction.

[40]We discuss the role and proper interpretation of egoism in section 5.5.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

	C	D
Cooperate (C)	\$2, \$2	\$0, \$3
Defect (D)	\$3, \$0	\$1, \$1

Figure 2. A version of prisoner's dilemma game form.

Hobbes's results, but these assumptions are much weaker than the assumption of psychological egoism.

5.2. The Bayesian game using PDGF

In this subsection, we describe how the PDGF provides the basis for a *bona fide* game, one that captures the incentives faced by heterogeneous agents in the state of nature.[41] In this game, agents in a large population randomly interact and face the material payoffs specified by the PDGF in Figure 2.[42] So, each player 1 is matched with a player 2. Each player role is indexed by $i \in \{1, 2\}$.

Although our model allows for an infinite variety of types, we follow Hobbes in classifying these types into two broad groups. First, there are what we call *dominators*. These types, according to Hobbes, take 'pleasure in contemplating their own power in the acts of conquest, which they pursue further than their own conservation requires' (Hobbes 1991: 88). In other words, these dominator types not only care about their relative standing, but also exhibit a strong desire to control more resources than others. Behaviourally speaking, they are willing to attack others in order to acquire greater power and a superior status, even when their survival does not depend upon it. To capture Hobbes's description of the dominator, the utility function must include a desire to possess more resources than others. There must be a preference for self-serving inequality. Let m_1 and m_2 be the material outcomes to player 1 and player 2 respectively. For example, if player 1 plays C and player 2 plays D, $m_1 = 0$ and $m_2 = 3$. A dominator type would then possess the following utility function:[43]

$$u_i^{\alpha_i}(m_i, m_j) = m_i - \alpha_i \max\{m_i - m_j, 0\} - |\alpha_i| \max\{m_j - m_i, 0\} \tag{1}$$

where $\alpha_i < 0$ is an exogenously given 'modesty' parameter of the payoff function. This utility function states that each citizen cares about her own material outcome, m_i but also cares about her material payoff in comparison to others. To see this, consider a particular utility function where $\alpha_i = -1$. Then, the second term dictates that, when player i has more than the other (i.e. $m_i - m_j > 0$), the second term becomes: $-(-1)(m_i - m_j) = (m_i - m_j) > 0$. In other words, the player gains utility from being ahead, just as Hobbes describes the dominator. In addition, the third term, $|\alpha_i| \max\{m_j - m_i, 0\}$, becomes: $|\alpha_i| \max\{m_j - m_i, 0\} = |\alpha_i| 0 = 0$. Adding

[41]The construction here is motivated by Sohn (2020) in that he models the prisoner's dilemma with heterogeneous social preferences.

[42]Since it is irrelevant who becomes player 1 and player 2 due to the symmetry of the PDGF, we assume that half of the citizens are playing the role of player 1 and the rest are playing the role of player 2.

[43]The proposed utility function is inspired by the inequity aversion model by Fehr and Schmidt (1999). While this class of utility functions may seem very special, we can show that our result does not require this class of utility functions. We show the extent to which our model can be generalized in section 5.5.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

up all the terms yields a payoff greater than the mere material payoff m_i. In general, whenever $\alpha_i < 0$, the player strives to be ahead and dislikes being behind. When behind (i.e. $m_i - m_j < 0$), only the third term becomes relevant since the second term becomes *0*. And the third term can only be (weakly) negative, which captures the disutility from being behind. The utility function with $\alpha_i < 0$ thus captures Hobbes's description of the dominator, and we therefore group any player with $\alpha_i < 0$ into the class of *dominator* types.

As we argued above, however, the majority of agents in the state of nature are *not* dominators. Instead, as Hobbes puts it, most 'would be glad to be at ease within modest bounds' (Hobbes 1991: 88). These individuals exhibit no intrinsic desire to be ahead. Instead, they merely wish to avoid being behind. Behaviourally, they will not attack others unless 'their own conservation requireth' that they do so (Hobbes 1991: 88). In particular, their utility functions will consist of two parts. First, because they care about their subsistence, they will always crave resources and power (Hobbes 1991: 70). Thus, one part of their utility function should be monotonically increasing in resource accumulation. Crucially, however, modest types will refrain from aggressive behaviour, even if this will yield a higher *material* payoff, so long as they have enough material resources to subsist. To capture this aspect of the *modest types*, a second part of the utility function must temper their desire to accumulate resources and power. Specifically, modest types have no desire to possess greater relative power than others. Their utility functions must therefore include a term that counteracts the desire for resources by decreasing insofar as their relative power becomes too great. Without such a term, the modest types would not 'be glad to be at ease within modest bounds'.

Interestingly, to model these modest types, we can use the same utility function as defined in (1), but it will look slightly different once we insert modest parameters. For the modest types, $\alpha \geq 0$, which means that such players prefer to avoid domination and subjugation. While the modest types would like more resources for self-preservation, they would not like to unilaterally defect when the co-player cooperates, though they still loathe being dominated by others. For modest types, i.e. those with $\alpha \geq 0$, we can simplify the utility function (1) to yield the following:

$$u_i^{\alpha_i}(m_i, m_j) = m_i - \alpha_i|m_i - m_j|.$$

For modest types, if the distribution of outcomes is unequal, then the agent experiences some 'disutility' (hence, the second term, $-\alpha_i|m_i - m_j|$). In other words, *ceteris paribus*, a modest type will not choose to possess more resources than her co-player, though, the acquisition of additional resources (m_i) may sometimes be worth the attendant inequality. At the same time, modest types do not wish to be behind others. Roughly speaking, one may say that they care about equality. And how much they care about equality is determined by the parameter, $\alpha_i \geq 0$. If $\alpha_i = 0$, the agent cares only about his material payoff and ignores relative standing. If $\alpha_i > 0$, on the other hand, the agent experiences disutility in proportion to the absolute difference of the material payoffs ($|m_i - m_j|$). This represents the modest types with a varying degree of concern for relative standing. That is, α_i is a measure of how vigorously the agent seeks to avoid inequality. This payoff function has the desired property that

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

	C	D
Cooperate (C)	$2, 2$	$-3\alpha_1, 3-\alpha_2$
Defect (D)	$3-\alpha_1, -3\alpha_2$	$1, 1$

Figure 3. PDGF with α_1 and α_2.

when the player is sufficiently high in modesty – i.e. α_i is large – universal cooperation is preferred to unilateral deviation. On the other hand, when the modesty parameter is sufficiently low, unilateral deviation is preferred to universal cooperation.

To illustrate the nature of the utility function, consider a preliminary case with two players, player 1 and player 2, who are matched to play the PDGF, and the values of the parameters, α_1 and α_2 are publicly known. Then, the game is described by Figure 3.

Note that mutual cooperation (C, C) is an equilibrium if and only if $\alpha_1, \alpha_2 \geq \frac{1}{3}$. In other words, if players are sufficiently modest (and if this fact is commonly known among players), then mutual cooperation is attainable.

Now, we study the game of interest, incorporating uncertainty. The timing of the game is as follows. First, each citizen of the population has the utility function as above, and each person's modesty parameter α_i is drawn from a commonly known, continuous cumulative distribution function[44] $F : (-\infty, \infty) \to [0, 1]$. Suppose that its support is an interval $[\underline{\alpha}, \overline{\alpha}]$. Thus, F determines heterogeneity of preferences in the population. Each citizen's value of α_i is private information; people cannot observe the other citizens' value of α_i. Then, each citizen is paired with another player to play the PDGF in Figure 2. Again, it is irrelevant for the theory who is assigned the role of player 1 and player 2.

Each citizen with a type α_i (or the α_i-agent) chooses to play either C or D. Without loss of generality, assume that the citizen only takes a pure strategy.[45] Then, all of the citizens' strategies can be summarized by a (measurable) decision function, $S : (-\infty, \infty) \to \{C, D\}$, so that $S(\alpha_i)$ dictates the strategy of the α_i-citizen. Now, we define the solution concept in the spirit of Bayesian Nash equilibrium.

Definition. *A decision function, $S : (-\infty, \infty) \to \{C, D\}$ prescribes an equilibrium if for every $\alpha_i \in (-\infty, \infty)$, the action $S(\alpha_i)$ maximizes the expected payoff for each α_i-citizen, given that the co-player cooperates with probability p_j, where p_j is the computed probability of j's cooperation from $S(\cdot)$. (Precisely, $p_j := \int_{-\infty}^{\infty} 1[S(\alpha_j) = C]dF$.)*

In other words, if everyone follows the action prescribed by the decision function S, then every citizen must be playing the optimal action also knowing that everyone else behaves according to the function, S. We have not assumed anything about F so far, but F is crucial as part of the analysis. In the later part of the analysis, we will assume more specific properties. But, for now, we shall discuss the general case, with

[44]By a *cumulative distribution function* (CDF) of a random variable X, we mean a function F that is defined as $F(x) := Pr(X \leq x)$. In other words, in our model a player's modesty parameter α is the random variable, and $F(\alpha_i)$ simply specifies the probability that a player i's utility function involves the modesty parameter α_i or any lesser modesty parameter. It can be shown that a CDF fully specifies a unique probability distribution over the values its random variable might take.

[45]This is without loss of generality in the sense that even if one allows for mixed strategies, no one ends up using a mixed strategy, except a measure-0 type, which is of no consequence in equilibrium.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

any arbitrary CDF, F. We will also elaborate on the extent to which F can be generalized (section 5.5).

5.3. Equilibrium characterization of the game

Before showing how this game and its equilibrium concept satisfy all the desiderata in Section 3, we must analyse the behaviour of individual agents and the properties exhibited by an equilibrium.

We begin by making an important observation that will allow us to both simplify our analysis and, as we argue later (section 5.4), to satisfy Desideratum 3. The observation is that any dominator type, i.e. any agent i with $\alpha_i < 0$, will play D, regardless of the co-player's action. This is intuitive, because playing D maximizes not only the material payoff, but also the psychological payoff. Since we know how all dominator types will behave, we only need to study the behaviour of modest types, i.e. those with $\alpha_i \geq 0$.

To begin this analysis, we must calculate the utility of a modest agent, which, of course, depends upon the probability that the other player will cooperate. Suppose that Player i believes Player j cooperates with probability p_j *ex ante*.[46] Given the value, p_j Player i's payoff from cooperation is given by

$$U_i^{\alpha_i}(C, p_j) := 2p_j + (1 - p_j)(-3\alpha_i).$$

Player i's payoff from defection is given by

$$U_i^{\alpha_i}(D, p_j) := 3(1 - \alpha_i)p_j + 1(1 - p_j).$$

Thus, Player i weakly prefers cooperation if and only if

$$2p_j + (1 - p_j)(-3\alpha_i) \geq 3(1 - \alpha_i)p_j + 1(1 - p_j). \tag{2}$$

By studying inequality (2), we make the following observation:[47]

Observation. (i) *Regardless of the value of α_i, no agent will cooperate if $p_j < \frac{1}{2}$.*

(ii) *For each agent with modesty parameter α_i there is a unique threshold $\tau(\alpha_i)$ such that C is the uniquely optimal strategy if and only if $p_j > \tau(\alpha_i)$.*[48]

(iii) *Any citizen whose type is less than $\frac{1}{3}$ never cooperates. Any citizen whose type is greater than or equal to $\frac{1}{3}$ is willing to cooperate, if others cooperate with a sufficiently high probability.*

The first part of the observation states that in order for any citizen to cooperate, the co-player *must* cooperate with a probability higher than $\frac{1}{2}$. The second part states that the minimum likelihood of cooperation one requires of the co-player (so that she would cooperate herself), depends on the value of α_i and is given by $\tau(\alpha_i)$. We shall call this value the *cooperation threshold* for α_i, or simply the *threshold* for α_i. This observation also implies that the higher α_i, the more lenient is player i, in the sense that the threshold level is lower. The third observation states that any citizen with $\alpha_i < \frac{1}{3}$ will never cooperate. These types will be called *unconditional defectors*, that is, players who will never cooperate regardless of the co-player's cooperation proba-

[46]By 'ex ante' probability of cooperation of j, we refer to the proportion of all citizens who play C.

[47]See appendix for technical details.

[48]$\tau(\alpha_i) := \frac{1+3\alpha_i}{6\alpha_i}$. Note that $\tau(\alpha_i)$ can be greater than 1.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

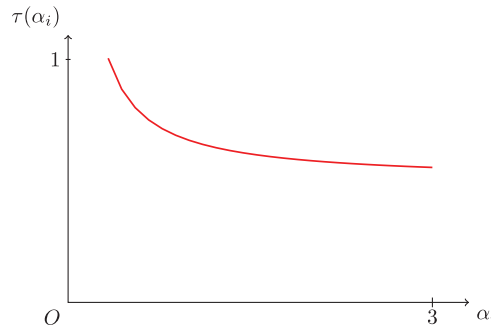

Figure 4. The mapping from types to thresholds.

bility. Clearly, the dominator types are unconditional defectors, but the converse is not true. And any citizen with $\alpha_i \geq \frac{1}{3}$ is willing to cooperate if the co-player cooperates with a sufficiently high probability. We shall call such types the *conditional cooperators*.

From Observation 1, we can deduce that *either* the α_i-citizen has a threshold value of $\tau(\alpha_i)$, *or* the α_i-citizen is an unconditional defector ($\alpha_i < \frac{1}{3}$). The mapping from α_i to the threshold for cooperation, $\tau : [\frac{1}{3}, \infty) \to [0, 1]$ is drawn in Figure 4. We will call this the *threshold mapping*.

We now have the technical apparatus required to find and characterize equilibria when the game has a general continuous CDF, F. After describing how to do so, we will compare equilibria in two cases: one where there are no unconditional defectors ($F \sim Unif[\frac{1}{3}, 1]$) and one that involves a small fraction ε of such defectors ($F \sim Unif[\frac{1}{3} - \varepsilon, 1]$).

There are two crucial facts that enable us to identify the equilibrium of this game. First, the proportion p of cooperators must be self-fulfilling. In other words, if p of the population are cooperating and the rest are defecting, then the cooperating citizens must not wish to deviate from their choice, and the same must be true for the defecting citizens.

Therefore, the equilibrium generally requires that for some value of $p \in [0, 1]$:

(i) p of the population are cooperating, and the rest are defecting.
(ii) Every player must be best-responding to the fact (*i*).

The second crucial fact is that if α_i-agent is cooperating, for any higher type, $\hat{\alpha}_i > \alpha_i$, $\hat{\alpha}_i$-agent must also be cooperating. Similarly, if α_i-agent is defecting, any lower type, $\underline{\alpha}_i < \alpha_i$, $\underline{\alpha}_i$-agent must also be defecting.

If we make use of these two facts, finding an equilibrium reduces to a simple procedure given by the following proposition:[49]

Proposition 1. *There exists an equilibrium where $p \in (0, 1)$ of the population cooperates if and only if there exists a value $\hat{\alpha}$ such that $\tau(\hat{\alpha}) = 1 - F(\hat{\alpha}) = p$.*

[49]Note that the corner cases are excluded in the proposition.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

The proposition states that in order for any cooperation to be feasible in equilibrium, we should be able to find a value $\hat{\alpha}$ that satisfies the equation, $\tau(\hat{\alpha}) = 1 - F(\hat{\alpha})$. Then, $1 - F(\hat{\alpha})$ determines the cooperation rate of the equilibrium.

The intuition for this result is as follows. By our first fact, the equilibrium level of cooperation p must make all agents happy with their choice of strategy. If it's an equilibrium where some people cooperate and some people defect, then, by continuity, there must exist someone in between, who is completely indifferent between cooperating and defecting. Call her $\hat{\alpha}$. By our second fact, it must be true that everyone with a modesty parameter above $\hat{\alpha}$ must be cooperating, and everyone with a modesty parameter below $\hat{\alpha}$ must be defecting. Therefore, the proportion of defectors in this equilibrium state is $P(\alpha_i \leq \hat{\alpha}) = F(\hat{\alpha})$, and the cooperation rate is $1 - F(\hat{\alpha})$. In short, to ensure that no one has incentives to deviate from this cooperation rate, $\hat{\alpha}$ must be indifferent. This will be true, if $\hat{\alpha}$ has a threshold $\tau(\hat{\alpha}) = 1 - F(\hat{\alpha})$. Satisfying this equality, therefore, ensures that no one wants to deviate.

Figure 5a depicts the threshold mapping $\tau(\alpha_i)$ in red alongside with a (complementary) CDF, $1 - F$. In this graph, the distribution is assumed to be uniform distribution on $[\frac{1}{3}, 1]$. According to Proposition 1, any cooperative equilibrium ($p > 0$) can be found by identifying points where the two curves coincide. In figure 5a, it is marked by a black dot at point A.

Point A shows that there is an equilibrium where all of the population cooperates. And this is achieved by the following state: have every type greater than or equal to $\frac{1}{3}$ cooperate and have the rest defect. But, the blue curve shows that everyone has a type above $\frac{1}{3}$. Therefore, the population achieves the cooperation rate of 1. At this state, no one wants to defect, since the lowest type in the population, $\frac{1}{3}$ is willing to cooperate, which implies that every type above $\frac{1}{3}$ is also willing to cooperate. So, this state is self-fulfilling, satisfying our definition of equilibrium.

Note that this graphical approach of finding intersections of the curves does not identify all equilibria. In particular, this method leaves out one equilibrium that always exists: the full defection equilibrium.

The fact that this full-defection equilibrium exists even when the two curves do not intersect will be crucial for understanding our derivation of war in the next example, so it's worthwhile to briefly examine this equilibrium. The rationale for the full defection equilibrium is quite intuitive: no type wants to cooperate when no one else cooperates. Suppose Player i believes that all of the potential co-players defect (i.e. $p_j = 0$). Then clearly, $U_i^{\alpha_i}(D, 0) > U_i^{\alpha_i}(C, 0)$ for all possible values of α_i. Thus the state of full defection will be self-fulfilling.

The fact that full defection is always an equilibrium shows that 'the state of war' is always a possibility, although this may not be the only possible outcome. Also note that if the lowest type, $\underline{\alpha}_i$ is greater than or equal to $\frac{1}{3}$, full cooperation is always an equilibrium, as shown in the example above. Essentially, the result we obtain in this case ($F \sim Unif[\frac{1}{3}, 1]$) is similar to the assurance game presented by Moehler (section 2): although the entire population consists of conditional cooperators, both full defection and full cooperation are possible equilibria.

So far we have discussed the method for finding equilibria and illustrated this method with a particular example where $F \sim Unif[\frac{1}{3}, 1]$. We now propose our reconstruction of Hobbes's argument, which begins by identifying a special class of CDFs. The only additional requirement that we impose on our CDF is that some

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

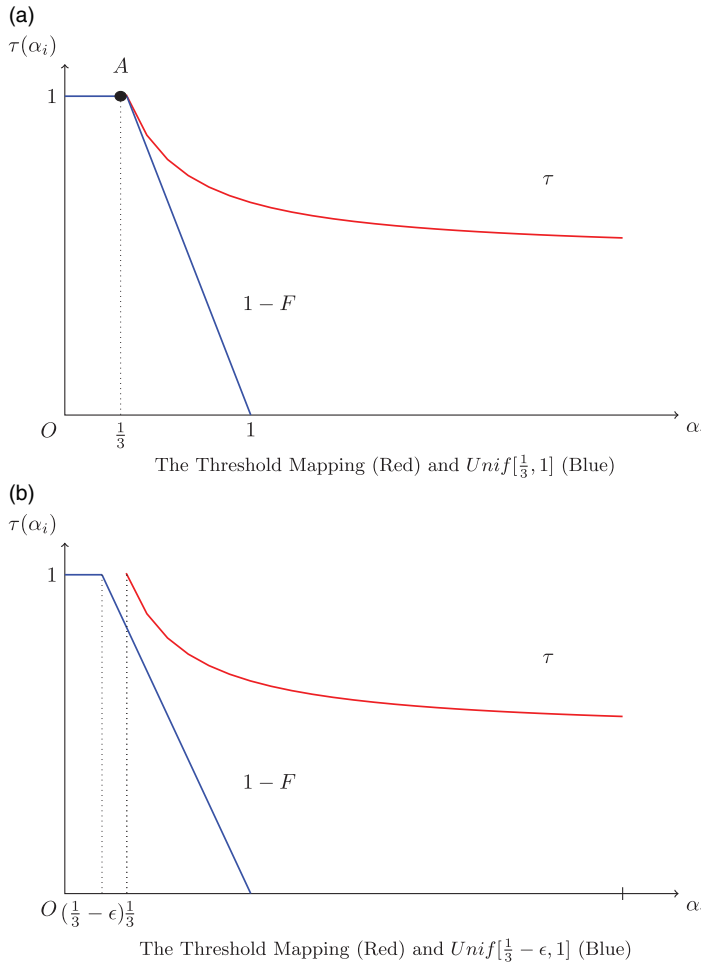

Figure 5. Threshold mappings and CDFs.

proportion, however tiny, of the population will defect. That is, they will choose the higher material payoff, regardless of how many others have chosen to cooperate. Given Hobbes's *arguably* egoistic view of human nature (Hobbes 1991: 102, 105, 109, 203), this is not a strong assumption.[50] We will demonstrate (section 5.4) that with this minimal assumption, our model satisfies all five desiderata. But first, let us consider how even a small proportion of unconditional defectors produces a trust-unravelling process that ultimately leads to war.

[50]In fact, in line with those who have denied that Hobbes endorses psychological egoism, this assumption only requires that *some* value material benefits *more* than they value cooperation and equity. And these individuals need not be psychological egoists. Rather, they are closer to what Kavka has labelled 'predominant egoists' (Kavka 1986: 64–80). We address the question of Hobbesian psychology in section 5.5.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

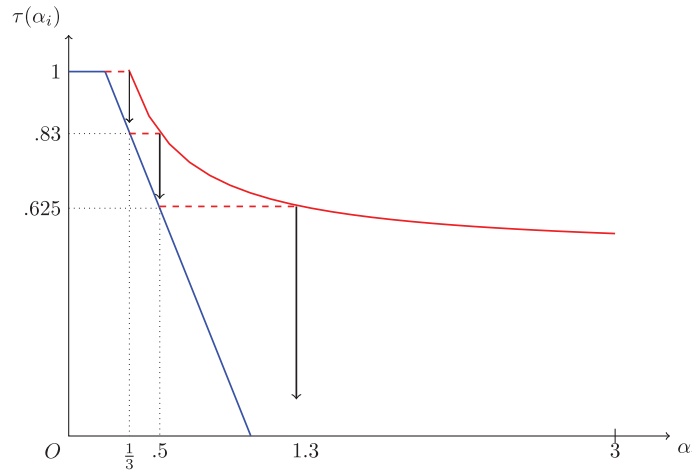

Figure 6. The threshold mapping (Red) and $1 - F(\cdot)$ (Blue).

With the addition of these unconditional defectors, we are now considering the uniform distribution on the support $[\frac{1}{3} - \epsilon, 1]$ as our choice of the CDF, F (see Figure 5b). Observe that the only difference between this distribution and the distribution in the earlier example is the existence of unconditional defectors. The following proposition shows that given this assumption on the CDF, the only equilibrium outcome is full defection. The state of war is inevitable.

Proposition 2. *For any value of* $\epsilon > 0$, *the only equilibrium is* $S(\alpha) = D$ *for all* $\alpha_i \in [\frac{1}{3} - \epsilon, 1]$.

Note that ϵ determines the proportion of unconditional defectors in the population, and the rest of the population is conditional cooperators, as shown in Figure 5b. So, the proposition shows that even if almost everyone in the population is a conditional cooperator, cooperation may still be impossible, due to the ϵ-presence of unconditional defectors. This is surprising, because when $\epsilon = 0$, full cooperation is possible. But, as soon as ϵ becomes strictly positive, there is no cooperation at all. This result also illustrates that we did not need a dominator type ($\alpha_i < 0$) at all to ensure the state of war. We merely need a tiny proportion of unconditional defectors ($\alpha_i < \frac{1}{3}$).

To understand how war emerges, refer to Figure 6. The blue curve depicts the complementary CDF ($1 - F$) of the uniform distribution on $[\frac{1}{3} - \epsilon, 1] = [.2, 1]$. Suppose that the population tries to support full cooperation in equilibrium. This initial state will be unstable, since common knowledge of rationality dictates that any type below $1/3$ must defect; these types never find cooperation optimal. Thus, it is common knowledge that the cooperation rate must be 0.83 at best. But then, since the cooperation rate is 0.83 at best, any type whose threshold is less than 0.83 ($\alpha < .5$) must also defect. Then, again, this fact must be common knowledge, if players recognize that others are also rational. This process of unravelling will continue until

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

no type is willing to cooperate. The elimination of cooperation is illustrated by the black arrows in Figure 6. Note that this reasoning happens in the minds of rational agents, not in their actual interactions. And the only epistemological requirement for this reasoning to hold is common knowledge of rationality. Even without actual experience with defectors or any process of learning, players of all types will recognize the imprudence of playing cooperatively. This iterative reasoning will thus eliminate cooperation completely. Only full defection survives as an equilibrium outcome.

There is a strong analogue between this phenomenon of unravelling and Akerlof's famous market-for-lemons model (Akerlof 1978). In both models, the driving force of failure is the asymmetry of information. In addition, the aggregate action of participants affect others' incentives to participate. However, the lemon market fails because a small number of lemons decrease the average quality of cars in the market, while, in our model, a small number of unconditional defectors diminish the average cooperation rate.

Time to take stock. This subsection has analysed the behaviour of agents that fit Hobbes's psychological description of the modest type. We have shown that when such agents interact in a structure determined by a prisoner's dilemma game form, those that are insufficiently 'modest' – that is, those with α_i below some value, $\frac{1}{3}$ in the example case – will always defect. These agents need not be *dominators* in the sense that they seek to dominate others. Rather, they are insufficiently modest to forgo material gain. When such agents are totally absent, full cooperation is possible. But when these unconditional defectors are present, even just an infinitesimal proportion of them, a process of *trust-unravelling* will ultimately lead all agents to choose defection. The model thus reveals how uncertainty drives the inevitable emergence of war, even in the absence of dominator types. Already, it seems clear that this new model will fare well with respect to our desiderata, but the next section endeavours to firmly establish this claim.

5.4. Satisfying the desiderata

Recall Desideratum 1.

Desideratum 1. The model should include a diversity of possible types. This means that, at a minimum, it should distinguish between conditional cooperators and unconditional defectors, and ideally it will include a wide spectrum of possible types, falling along a continuum from altruistic cooperator to egoistic aggressor.

It is clear that this desideratum is satisfied in our model, as it admits a continuum of types via the cumulative distribution function, F.

Now, recall Desideratum 2.

Desideratum 2. Players should be uncertain as to the type (viz. payoff function) of their opponents, and this should make a significant difference to their choice of strategy.

By assumption of private information, players are uncertain as to the type of opponents. We argue that the prediction of the model is significantly different. In order to show that the uncertainty in fact makes a significantly different prediction, we must have a counterfactual game where there is no uncertainty, and make comparative statics among the two environments. So, we consider the

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

identical game except that players can observe opponents' types, and thus there is no uncertainty.

Supposing for the moment that players *can* observe the other citizens' types, then, we can show that mutual cooperation is an equilibrium outcome of a given pair of citizens as long as both citizens' types are greater than or equal to $\frac{1}{3}$. Therefore, we can easily compute the potential for cooperation without uncertainty. The proportion of pairs with both players' types weakly greater than $\frac{1}{3}$ is $\left(\frac{1-\frac{1}{3}}{1-\frac{1}{3}+\epsilon}\right)^2 = \left(\frac{2}{2+3\epsilon}\right)^2$.

So, depending on the value of ϵ, in the game without uncertainty, the mutual cooperation rate in equilibrium is at least $\frac{4}{9}$. With a small value of ϵ, the population can achieve an arbitrarily high rate of cooperation. This statement is formalized in the observation below.

Observation. *Choose any large q such that* $0 \leq q < 1$. *If* ε *is sufficiently small, there exists an equilibrium where the population achieves a mutual cooperation rate (i.e. the proportion of pairs achieving* $[C, C]$ *in equilibrium) greater than q.*

This observation serves to show that uncertainty does, in fact, lead to a significantly different prediction when compared with the game without uncertainty. Without uncertainty, any arbitrarily high cooperation rate is possible if ϵ is small enough. But under uncertainty, cooperation is impossible for *any* ϵ. Thus, Desideratum 2 is satisfied.[51]

The fact that cooperation is impossible for *any* ϵ, even infinitesimal values, points towards one way in which the Trust-unravelling Model might satisfy the third desideratum.

Desideratum 3. The model should demonstrate the emergence of war without relying on the presence of dominator types.

Although it is true that our model improves upon several past models by relying only on an infinitesimal proportion of unconditional defectors, this would not allow us to claim that it fully satisfies Desideratum 3, nor is it the argument that we have made.[52] Rather, we have carefully distinguished between dominator types and unconditional defectors. Though all dominator types are unconditional defectors, some modest types, viz. some agents with $\alpha_i > 0$, may also be unconditional defectors. Thus, there is a class of modest types that are, nonetheless, unconditional defectors. Call these unconditionally defecting modest types *marginally modest types*. The model also demonstrates how, under conditions of uncertainty, a tiny proportion of marginally modest types will initially defect, causing a chain reaction that ultimately leads to full defection.[53] These marginally modest types defect, not out of any intrinsic desire to domineer over others, but simply because their preference for equity is too weak to outweigh the potential material gains they can

[51]More precisely, let any $q \in (0, 1)$ be the difference in mutual cooperation rate between the two environments required to call the two 'significantly different'. We can always find a game that passes this 'significant difference' test, for any $q \in (0, 1)$, by choosing ϵ sufficiently small.

[52]Though we do revisit this argument at the end of section 6.

[53]Again, this 'chain reaction' is a process of reasoning that occurs in the minds of agents deciding whether to cooperate or defect.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

accrue by defecting. Again, even though this proportion of unconditional defectors, ϵ, may be infinitesimal, the population will fail to achieve any mutual cooperation. This renders our model consistent with an interpretation of Hobbes in which dominator types play a non-necessary role in fomenting war. Of course, if dominator types are highly prevalent, our model will still predict war. However, the above observation has allowed us to demonstrate that even if no citizens find intrinsic satisfaction in 'acts of conquest' (Hobbes 1991: 88), war will still emerge. The model thus satisfies the third desideratum.

There is an important objection here concerning our terminology. We do not equate dominators with unconditional defectors as other commentators have. Consequently, when we claim to satisfy this desideratum in contrast to other models that fail to satisfy it, e.g. the prisoner's dilemma, we appear to be applying a double standard. We will address this objection more thoroughly at the end of section 6. For now, it's important to note that our distinction between dominators and unconditional defectors is possible only because we have introduced distinctively social preferences that better capture Hobbesian psychology than past models. Other models, therefore, have little to say about whether or not dominators (in our strict sense) are necessary for the emergence of war, since they do not explicitly identify which of their defectors are truly dominators who enjoy domineering over others and which are merely interested in material gain. Our model is uniquely able to explicitly distinguish these two types and to thereby show that war does not require *bona fide* dominators.

Finally, Desideratum 4 and Desideratum 5 are satisfied quite naturally. Under conditions of uncertainty, the only possible equilibrium is that of full defection, which represents the state of war. We have shown this in a purely deductive manner, without relying on empirical observation of computer-based simulations.

Desideratum 4. The model should demonstrate that war is the unique equilibrium of the model.

Desideratum 5. The model should derive its equilibrium in a purely deductive, a prior manner, as Hobbes intended by his geometric method.

Through this discussion, we have informally demonstrated the key result of this paper.[54]

Theorem. *The game, along with the solution concept, satisfy Desiderata* 1–5.[55]

5.5. Comments on the generality of the model

In the model, we have, for now, assumed a specific structure on the material payoffs of the PDGF, and the class of payoff functions we consider. One may wonder how general the model can be made. In other words, one may wonder whether the results we provide depend on the specific structure of the PDGF and the payoff functions. Surprisingly, the PDGF and the class of utility functions can be made very general as long as it has the prisoner's dilemma structure in Figure 7.

Then, two additional assumptions suffice to yield our result: (1) players are expected-payoff maximizers, a standard game-theoretic assumption, and (2) each

[54]A more formal demonstration can be found in the supplementary appendix.

[55]For some choice of $\epsilon > 0$.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

	C	D
Cooperate (C)	$\$c, \$c$	$\$0, \$x$
Defect (D)	$\$x, \$0$	$\$d, \$d$

Figure 7. The general PDGF ($x > c > d > 0$).

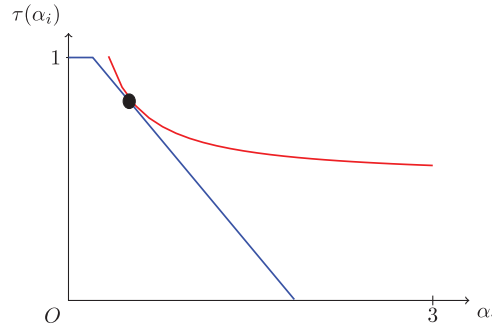

Figure 8. Partially cooperative equilibrium.

citizen prefers defection when the co-player defects for sure. Our result requires no further assumptions. Note that for the utility function we proposed in the previous subsection, for every value of α_i, the agent preferred defection when the co-player defected (because $u_i^{\alpha_i}(d,d) > u_i^{\alpha_i}(0,x)$ for every α_i).[56] The Trust-unravelling Model therefore predicts war without placing strict constraints on the utility functions of modelled agents.

The constraints placed on the *distribution* of these utility functions (i.e. the prevalence of various levels of modesty parameter α) are also fairly weak. In order to derive war as the unique result, we have assumed a special class of population distributions which may seem to limit the generality of the model, but this class is quite broad. For our main theorem to hold, there is one crucial assumption, illustrated in Figure 6. As Figure 6 shows, whenever the blue curve $(1 - F)$ is below the red curve (the threshold mapping), unravelling will inevitably occur, leaving the state of war the only possibility. And if the blue curve intersects with the red curve at any point (see Figure 8), the unravelling may not eliminate cooperation completely, as there is an intermediate equilibrium where some of the population actually cooperate.

What is the philosophical-psychological meaning of this graphical configuration? Stating its precise meaning in non-mathematical terms is difficult, but it roughly means that the population is not too altruistic. This is an interesting result, since it bears directly on a long-standing debate regarding the nature and extent of Hobbes's commitment to psychological egoism. Many commentators have taken Hobbes to expressly defend an egoistic psychological theory, in which all human

[56]This fact has been first shown by Sohn (2020). For readers interested in full technical details of the argument made here, see Sohn (2020).

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

beings are always and everywhere driven by self-interested ends.[57] Others have claimed that such an interpretation ignores important passages and renders Hobbes's overall system internally inconsistent.[58] Surveying the evidence with apparent consternation, Kavka writes that 'no systematic pattern of development is discernible here; apparently Hobbes appeals to whichever version of egoism seems best to support the point he is trying to make at the time' (Kavka 1986: 45).

Kavka goes on to defend an interpretation that will save Hobbes's argument without committing Hobbes to strict psychological egoism, calling this new view *predominant egoism* (Kavka 1986: 64–65). Predominant egoism states that 'self-interested motives tend to take precedence over non-self-interested motives in determining human action' (Kavka 1986: 64). Kavka's approach is promising – in trying to interpret Hobbes's ambiguous statements on human psychology, it *does* make sense to seek out the most plausible form of egoism that will sustain the validity of Hobbes's arguments. However, saying that 'self-interested motives tend to take precedence over non-self-interested motives' is itself somewhat ambiguous. The assumption we place on our population distribution provides a plausible way of filling in the details. In effect, we have shown one way to make the theory of predominant egoism more precise. Although Hobbes did not possess the tools of modern decision theory, he had a hunch that egoism of some form plays an important role in the emergence of war. If our reconstruction of his argument is on track, then we have uncovered a precise and fairly weak psychological assumption that underlies Hobbes's vision of trust-unravelling in the state of nature. In other words, if our model successfully captures Hobbes's reasoning, then we have rationalized and clarified Hobbes's belief that some form of egoism is necessary to generate his result.

6. Summary and clarification

The table in Figure 9 summarizes the status of all the models considered so far with respect to the five desiderata.

	Desideratum 1	Desideratum 2	Desideratum 3	Desideratum 4	Desideratum 5
Prisoners' Dilemma	X	X	X	O	O
Assurance Game	X	Δ	O	X	X
Chung's Model	Δ	O	Δ	O	O
Vanderschraaf's Model	O	O	Δ	O	X
Trust-unravelling Model	O	O	O	O	O

X = Fails to satisfy

Δ = Partially satisfies

O = Fully satisfies

Figure 9. Summary table.

[57]Advocates of this interpretation include Broad (1949), Watkins (1965), Gauthier (1969) and Butler (2017).

[58]For this view, see Gert (1965, 1967), McNeilly (1966), or Chung (2016). Barrett (2020) has also recently expressed concurrence with this idea.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

As mentioned above, one might object to our handling of Desideratum 3. In short, in declaring the Trust-unravelling Model to satisfy it while other models fail to, we have implicitly assumed two different definitions of 'dominator type'. One definition, that adopted by Chung and Vanderschraaf, is that a dominator type is simply an agent that will choose to defect even when the co-player chooses to cooperate. The other definition, which we employ here, takes a dominator type to be one who takes joy in self-serving inequality. As above, let us distinguish between these two types. The first are what we have called *marginally modest agents*, or those with $\alpha_i > 0$ who nevertheless unconditionally defect. These types are not, according to our categorization, dominator types, but to avoid biased terminology we shall refer to them as 'dominator* types'. The second notion of dominator, what we hold to be the *bona fide* dominator type, includes any agent who takes joy in self-serving inequality. We continue to refer to this second type of dominator as, simply, a *dominator*. Now, the objection goes, by showing that our model does not rely on dominator types, we have not shown that it doesn't rely on dominator* types.

This last claim is, of course, correct. However, so is the following claim: past models have not shown that they do not rely on dominator types. In fact, past models lack the resources to distinguish between dominator* and dominator types, because they do not incorporate Hobbesian psychology in the way that the Trust-unravelling Model does. By including social preferences as a basis for agents' decision-making, we can distinguish between different reasons for defecting. Hence, if we consistently apply the second definition and consider only *bona fide* dominator types, then the Trust-unravelling model is the only one with the resources to demonstrate that such types are not required for war to emerge.

Suppose, on the other hand, that the objector insists dominator* types are *bona fide* dominator types. The objector may assert that any unconditional defector is properly characterized as a dominator type, regardless of whether the agent takes joy in self-serving inequality. Accepting this broader definition, we might distinguish three levels of satisfying Desideratum 3:

Level 1 (weak): Even an arbitrarily small proportion of unconditional cooperators will drive the system towards war.

Level 2: Even an arbitrarily small proportion of unconditional cooperators will drive the system towards war, even if no one actually interacts with an unconditional cooperator.

Level 3 (strong): The system will be driven to war whether or not there are any unconditional cooperators present at all.

According to these levels, Vanderschraaf's model outperforms Chung's by placing less importance on dominator types: only a vanishingly small proportion of dominator types need be present. For this reason, Vanderschraaf's model satisfies the first level of Desideratum 3. However, the Trust-unravelling Model goes further: players need not even encounter a dominator type before determining that anticipation (i.e. defection) is their best response. The process of trust-unravelling is entirely rational, not experiential. Neither model, however, succeeds in satisfying the third level of Desideratum 3, the complete absence of

all unconditional defectors from the population. Only Moehler's assurance game, disqualified for other reasons, attains this honour.[59]

In sum, whether we adopt the notion of dominator or dominator*, the Trust-unravelling Model outperforms the two most impressive models to date, Chung's and Vanderschraaf's. Desideratum 3 thus supports the claim that this model makes progress in more accurately reconstructing Hobbes's argument in *Leviathan*.

In sum, the Trust-unravelling Model compares quite favourably with both second wave models. While Chung's model allows for only two types of player, the Trust-unravelling Model allows for an infinite array of distinct utility functions. All three of these models satisfy Desideratum 2, viz. that uncertainty should play an important role in producing war. This is the key fact that makes all of these models part of the second wave. What distinguishes the most cutting edge state of nature models from less advanced approaches is that they take seriously Hobbes's insistence that trust, fear or uncertainty play a key role in undermining cooperation. Finally, while Vanderschraaf's model fails to satisfy Desideratum 5 due to its reliance on simulations rather than proofs, the Trust-unravelling Model, like Chung's model, is purely a priori, deriving its results with absolute certainty. It thus captures Hobbes's geometrical aspirations.

7. Conclusion

The Trust-unravelling Model incorporates a collection of important aspects of Hobbes's argument that war must emerge from the state of nature. These aspects, to reiterate, are that individuals in the state of nature should exhibit diverse preferences, that their uncertainty should play a crucial role in determining their choice to preemptively attack, that the argument should rely as minimally as possible on aggressive types, and, finally, that the argument should follow Hobbes's methodology, showing deductively that war will emerge from the state of nature. As we have seen, past models – including the impressive models of the second wave – do not capture all five of these aspects. The Trust-unravelling Model thus emerges as an improved reconstruction of one of Hobbes's central arguments in *Leviathan*.

As a parting thought, consider how the Trust-unravelling Model has revitalized the classic prisoner's dilemma. Although we have rejected the prisoner's dilemma as an adequate model, we have nevertheless shown it to be a useful tool for understanding Hobbes's argument. The key lies in the distinction between a *game* and a *game form.* The prisoner's dilemma *game* deviates importantly from Hobbes's argument, yet we have embraced the prisoner's dilemma *game form*. One can accept the idea that resource units might be determined roughly in accord with a prisoner's dilemma game form while rejecting the idea that preferences are necessarily monotonic in resource units. The nature of the rules underlying resource allocation may explain why so many have seen Hobbes's state of nature as intuitively presenting a prisoner's dilemma. When Rawls claimed that Hobbes's state of nature furnishes the 'classical example' of the prisoner's dilemma (Rawls 1999: 238), he may have been more on-target than it seems. The error arises from confusing subjective

[59]Perhaps future work will explore ways to achieve this level without exhibiting the defects of Moehler's model, or else demonstrate the impossibility of deriving universal defection as a unique equilibrium in the total absence of unconditional defectors.

https://doi.org/10.1017/S0266267121000079 Published online by Cambridge University Press

payoffs with material resource units. Rather than addressing this erroneous assumption, past attempts to salvage the prisoner's dilemma have turned to repeated versions of the game. But this has produced little improvement.

In fact, this observation may be useful for other political theorists participating in the research programme of reconstructing past theories using game-theoretic techniques. We suspect that many games that scholars have employed may be too simple for deeply exploring most political theories. In particular, simple models do not adequately capture important features of human nature that political theorists have considered important to deriving their results. However, when interpreted as a *game form*, and appended with a more psychologically rich payoff function, it is possible to capture certain features that must otherwise be left out. Our model illustrates this possibility. In the Trust-unravelling Model, this approach helped in roughly two ways. First, it allowed us to model citizens as exhibiting a more plausible psychology, and second, it allowed us to incorporate the inevitable uncertainty regarding the psychology of others. We hope that other theorists find this general approach helpful in tackling novel problems.

Acknowledgements. We would like to thank the participants of the 2020 PPE Society Annual Meeting, who provided excellent comments on a draft of this paper, as well as two anonymous referees at *Economics and Philosophy* who suggested many worthwhile revisions. Alexander would also like to thank the students in his Spring 2019 class, 'The Social Contract', who patiently corrected his errors and motivated him to study Hobbes more thoroughly than he had ever hoped to.

Supplementary material. To view supplementary material for this article, please visit: https://doi.org/10.1017/S0266267121000079

References

Akerlof G.A. 1978. The market for 'lemons': quality uncertainty and the market mechanism. In *Uncertainty in Economics*, 235–251. New York, NY: Academic Press.

Alexandra A. 1992. Should Hobbes's state of nature be represented as a prisoner's dilemma? *Southern Journal of Philosophy* **30**(2), 1–16.

Barrett J. 2020. Punishment and disagreement in the state of nature. *Economics and Philosophy* **36**, 1–21.

Binmore K. 2005. *Natural Justice*. New York, NY: Oxford University Press.

Broad C. 1949. Egoism as a theory of human motives. *Hibbert Journal* **48**, 105–114.

Bruner J.P. 2015. Diversity, tolerance, and the social contract. *Politics, Philosophy & Economics* **14**, 429–448.

Butler J. 2017. *Fifteen Sermons Preached at the Rolls Chapel: And Other Writings on Ethics*. New York, NY: Oxford University Press.

Chung H. 2015. Hobbes's state of nature: a modern Bayesian game-theoretic reconstruction. *Journal of the American Philosophical Association* **1**, 485–508.

Chung H. 2016. Psychological egoism and hobbes. *Filozofia* **71**, 197–208.

Chung H. 2018. Rawls's self-defeat: a formal analysis. *Erkenntnis* **85**, 1–29.

Cooper R., D.V. DeJong, R. Forsythe and T.W. Ross 1992. Communication in coordination games. *Quarterly Journal of Economics* **107**, 739–771.

Delmotte C. 2020. Tax uniformity as a requirement of justice. *Canadian Journal of Law and Jurisprudence* **33**, 59–83.

Fehr E. and K.M. Schmidt 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* **114**, 817–868.

Gauthier D.P. 1969. *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*. New York, NY: Oxford University Press.

Gert B. 1965. Hobbes, mechanism, and egoism. *Philosophical Quarterly* **15**, 341–349.

Gert B. 1967. Hobbes and psychological egoism. *Journal of the History of Ideas* **28**, 503–520.

Hampton J. 1988. *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press.
Harsanyi J.C. and R. Selten 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.
Hausman D.M. 1992. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.
Hausman D.M. 2011. *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.
Henrich J. 2017. *The Secret of our Success: How Culture is Driving Human Evolution, Domesticating our Species, and Making us Smarter*. Princeton, NJ: Princeton University Press.
Hobbes T. 1991. *Leviathan*. Cambridge: Cambridge University Press.
Hobbes T. 1998. *On the Citizen*. Cambridge: Cambridge University Press.
Jesseph D. 1996. Hobbes and the method of natural science. In *The Cambridge Companion to Hobbes*, 86–107. Cambridge: Cambridge University Press.
Kandori M., G.J. Mailath and R. Rob 1993. Learning, mutation, and long run equilibria in games. *Econometrica* **61**, 29–56.
Kavka G.S. 1983. Hobbes's war of all against all. *Ethics* **93**, 291–310.
Kavka G.S. 1986. *Hobbesian Moral and Political Theory*. Princeton, NJ: Princeton University Press.
Kavka G.S. 1989. Political contractarianism. Unpublished manuscript.
Kogelmann B. and B.G. Ogden 2018. Enough and as good: a formal model of Lockean first appropriation. *American Journal of Political Science* **62**, 682–694.
Kogelmann B. and S.G. Stich 2016. When public reason fails us: convergence discourse as blood oath. *American Political Science Review* **110**, 717–730.
Lehtinen A. 2011. The revealed-preference interpretation of payoffs in game theory. *Homo Oeconomicus* **28**, 265–296.
McNeilly F. 1966. Egoism in Hobbes. *Philosophical Quarterly* **16**, 193–206.
Moehler M. 2009. Why Hobbes' state of nature is best modeled by an assurance game. *Utilitas* **21**, 297–326.
Moehler M. 2020. *Contractarianism*. Cambridge: Cambridge University Press.
Ostrom E. and J. Walker 2000. *Polycentric Games and institutions: Readings from the Workshop in Political Theory and Policy Analysis*. Ann Abor, MI: University of Michigan Press.
Rawls J. 1999. *A Theory of Justice*: Revised Edition. Cambridge, MA: Harvard University Press.
Ristroph A. 2014. Hobbes on 'diffidence' and the criminal law. *Foundational Texts in Modern Criminal Law* **23**, 31.
Skyrms B. 2001. The stag hunt. *Proceedings and Addresses of the American Philosophical Association* **75**, 31–41.
Sohn J. 2020. Cooperation with uncertain social preferences. Working Paper.
Thrasher J. and K. Vallier 2018. Political stability in the open society. *American Journal of Political Science* **62**, 398–409.
Turner P.N. and G. Gaus 2017. *Public Reason in Political Philosophy: Classic Sources and Contemporary Commentaries*. London: Routledge.
Van Huyck J.B., J.P. Cook and R.C. Battalio 1997. Adaptive behavior and coordination failure. *Journal of Economic Behavior & Organization* **32**, 483–503.
Vanderschraaf P. 2006. War or peace? A dynamical analysis of anarchy. *Economics and Philosophy* **22**, 243–279.
Watkins J.W. 1965. *Hobbes's System of Ideas*. London: Hutchinson & Co.

Alexander Schaefer is a PhD student in philosophy at The University of Arizona and a Politics, Philosophy, Economics and Law Fellow at the Freedom Center. His research interests include social contract theory, political economy and social complexity. His current projects focus on applying the insights of complexity theory to assess the proper scope of state action. Email: schaefer1@email.arizona.edu. URL: alexanderschaefer.weebly.com

Jin-yeong Sohn is an Assistant Professor of Economics at the Institute for Advanced Economic Research, Dongbei University of Finance and Economics. His research interests include behavioural economics, experimental economics and game theory. In particular, he is interested in reciprocity theory, social preferences and psychological game theory. URL: http://iaer.dufe.edu.cn/html/People/faculty/2020/1013/21.html

Cite this article: Schaefer A and Sohn J (2022). Unravelling into war: trust and social preferences in Hobbes's state of nature. *Economics and Philosophy* **38**, 171–205. https://doi.org/10.1017/S0266267121000079