Evolutionary Models and the Normative Significance of Stability[1]

Arnon Levy
The Hebrew University of Jerusalem
Arnon.levy@mail.huji.ac.il

Forthcoming in *Biology & Philosophy*

1. Introduction

Many have expected that understanding the evolution of norms should, in some way, bear on our first order normative outlook: How norms evolve should shape which norms we accept. But to date, there has not been much to shore up this expectation. Most existing discussions of evolution and norms take one or the other of the following forms. On the one hand, there are those who attempt to "overhaul" moral and political thinking on the basis of accounts of their evolutionary underpinnings. On the other hand, some use evolutionary considerations to argue for theses in meta-ethics, such as various forms of anti-realism. Both kinds of discussion carry considerable interest, but can also become frustrating. The first tends to jump headlong into the is/ought gap – proclaiming that one can directly derive one's favored moral principles from a descriptive, often rather speculative, evolutionary story (extreme cases in point: Ruse, 1986; Wilson & Ruse, 1986). The second kind of project has seen a much more sophisticated development by such authors as Street (2006) and Joyce (2006) and concomitantly subtle responses (Fitzpatrick, 2014 provides an overview). But it tends to abstract almost entirely from the details of evolutionary theory, and may effectively be a version of an important yet well-known problem in moral epistemology (Enoch, 2011).

My aim in this paper is to describe a third option – another way in which evolutionary considerations can feed into normative thinking. In particular, I will discuss two related forms of argument that utilize information about social stability drawn from evolutionary models and employ it to assess claims in first-order political philosophy. The kind of arguments I have in mind aren't new. Indeed, they have precedents going back at least to early modern times (some of which are noted below). In a nutshell, one form of argument treats stability as feature of social states that

[1] For comments on previous versions of this article I am grateful to Christine Clavien, Ittay Nissan-Rozen, David Wiens.

can (and ought to) be taken into account alongside other features. Another form of argument views stability as a constraint on the realization of social ideals, via a version of the ought-implies-can maxim. As I say, these kinds of arguments have a history in political philosophy. But their marriage with evolutionary information is relatively recent, has a significantly novel character, and has received scant attention in recent moral and political philosophy. In light of this, I will primarily aim to explain how such arguments work; but I will also, to an extent, evaluate their potential merits.

A couple of remarks should help set the stage and clarify what is to come. First, I will presuppose a cultural interpretation of evolutionary models. Cultural evolution, as I use the term here, is structurally similar to familiar biological, Darwinian evolution – it is a process in which those who have traits that enhance their chances of survival and reproduction spread. But rather than traits being transmitted biologically, via genes and sex, in cultural evolution traits are transmitted socially, through learning from others. The emerging dynamics are not identical to those of biological evolution by natural selection (Lewens, 2015), but these differences will not matter for present purposes.[2] Regarding the evolution of norms as cultural is consonant with how many modelers in this area view their work. They view it so, among other reasons, due to the relative speed at which norms change, and due to the variation in norms across different societies. That said, I think that key parts of the discussion can be extended to biological evolution. But I will not do so here, both because the extensions are fairly straightforward and because that would take up too much space.

A second remark concerns the temporal aspect of evolutionary models. Descriptive work on evolution, including the evolution of norms, is usually taken as retrospective in character. The aim is to describe a process that has already occurred, in order to explain extant phenomena. But some theoretical work in this area, especially the more abstract mathematical parts, can be read prospectively too – or modally, if you will – as telling us about future and/or possible states of affairs. Put differently: in this mode, one is regarding the models as predictive rather than backward-looking explanations. The predictive\prospective reading of models is necessary for the kind of arguments I want to explore. But as we shall see, it introduces some complications into their application and into the assessment their merits.

The paper attempts to combine ideas from evolutionary theory, political philosophy and philosophy of science, roughly in that order. First, in section 2, I will describe some work by Brian

[2] In some cases, cultural and biological evolutionary models are formally equivalent. But that depends on particular assumptions (Weibull, 1995, Ch. 4). In general, the processes are similar but distinct.

Skyrms and his collaborators, as an illustration of the kind of evolutionary work I have in mind. Section 3 is devoted to the notion of stability – I provide a characterization of stability, as it pertains to social states, outline two ways in which it comes into play in political thinking, and situate this form of argument within some relevant discussions of stability in political philosophy. The resulting arguments from stability are conceptually uncomplicated, I think. But issues arise when we look to put such arguments into practice, and these will be discussed in section 5. Section 6 provides a summary.

2. Skyrms and Alexander on Divide the Dollar

The example I'll describe in this section, and use throughout, is a model of the evolution of equal sharing developed by Brain Skyrms and several collaborators (Skyrms 1996, 2003; Alexander, 2007; Harms and Skyrms, 2008 provide an overview). The model is a classic in this area, worth focusing on in and of itself. But it is also a clear illustration of an important class of models that are offered as accounts for the origins of norms. I'll first describe the model and the results obtained by Skyrms et al., discuss its structure – as that is the main aspect I wish to illustrate – and then, relatively briefly, address the extent to which key features of this example generalize.

Skyrms and several collaborators, primarily Jason McKenzie Alexander, used the well-known game of divide the dollar (AKA the Nash bargaining game). In this game, two individuals make independent demands for a share of a unit good, the cake. If demands sum to 100% or less, each gets what they demanded. Otherwise both get nothing. In particular, if both players make a demand for half the cake then the good is split equally – so that they are effectively employing a "share and share alike" rule, a rudimentary egalitarian principle.

Skyrms and Alexander constructed an evolutionary version of divide the dollar: there is a population of players, each with a strategy corresponding to a demand, such as 50%, 60% or 40%. The model was developed over time – it would be more accurate to think of it as a family of models – but in its mature form the population is organized as a network, specifically a square lattice. The game is played over multiple rounds. In each round, each player is paired with every individual in its immediate neighborhood (thus, the game is still pair-wise). Subsequent to each such round, players update their strategy. In one key version of the model, this is done via the rule "imitate your best neighbor" (IBN): each player surveys the neighbors and updates their strategy to match the strategy of their most profitable neighbor. In each run of the model, then, this game is played over and over again (typically, in computer simulations, for tens or hundreds of thousands of rounds), and modelers observe the changing distribution of strategies in the overall population.

As noted, this version of the model makes certain assumptions—I'll refer to these as *structural assumptions*—specifically, the network is assumed to be a regular lattice, and updating is done via IBN. These elements of the model can be varied: different network structures can be employed and rules other than imitation can be chosen. Indeed, Alexander (2007) is a book length treatment of such models. In it, he looks at various versions of divide the dollar, including ones that include different kinds of networks (such as small world type networks, and networks that change over time). He also tests different updating rules, including other forms of imitation (e.g. 'imitate proportional to success'; 'imitate the best in the overall population'). In principle, other update rules are possible, including "smarter" ones, such as various forms of hypothesis testing, best-response (which attempts to predict the neighbor's strategy, and maximize payoff relative to it), or "dumber" ones, such as reinforcement learning (Young, 2004).

Back to the version studied by Skyrms and Alexander: They ran simulations in which the initial composition of the population was varied, and found that from almost every initial state, a population in this game eventually reaches a state in which demand ½ dominates. This is a demonstration that, at least in this kind of setting, equal splitting is a stable outcome. If one moves the population to an unequal split – i.e. to a distribution of strategies significantly different from all-demand-½ – it is highly likely to return to equal splitting.

There is much to learn from this example in its own right, and many questions may be raised about what it is meant to model – an egalitarian rule of conduct, or merely a set of behaviors, for example; and whether the modeling assumptions, for instance the IBN update rule, are plausible (Levy, 2011). Here, however, my aim is to discuss its basic from, which illustrates a broad class of models.[3] It involves four elements:

(1) A game like divide-the-dollar, or more generally an interpersonal interaction such as a cooperation/defection scenario or a bargaining game. The game represents a social problem that the norm responds to. Typically, there is a specific solution of interest, such as a 50:50 split or mutual cooperation.

(2) A population of interactants; and a distribution of strategies on it. Typically, as noted above, a simulation study of such a model includes "runs" from a range of different initial populations, and the distribution of strategies is what evolves over time.

[3] Alexander (2007) presents similar models for norms of cooperation (Prisoner's Dilemma), trust (Stag Hunt), retribution (Ultimatum), for instance.

(3) An interaction structure which determines how and when members of the population interact, such as a network with neighborhoods defined on it, and a specification of the pairing mechanism, for instance sequential pairing with neighbors (in principle, pairing could be random).

(4) Lastly, an update rule which determines how players change strategies over rounds of the game – for instance through IBN (though, as noted, a range of other rules can be employed.)

The goal of many models in this area, as with the Skyrms-Alexander model, is to show that the solution of interest is (or isn't) stable, i.e. that the population will arrive at it from a variety of starting points, sometimes also under a range of assumptions about the interaction structure and/or the update rule. (Below I will say more about stability, in these models and in general). Skyrms and Alexander do this via computer simulations; in some cases, one can prove a stability result analytically. The basic goal is similar.[4]

Now, the typical use of such model is explanatory and retrospective: the goal is to explain an existing phenomenon, such as a tendency towards equal sharing.[5] The usual mode in which such theorizing is done involves a heavy focus on stability, for what are, broadly speaking, epistemic reasons. It is difficult to obtain information about processes in the past—a general impediment in studying evolution—and this includes both information about initial conditions as well as about structural aspects. However, if a particular state can be shown to be stable then the lack of knowledge of initial conditions and specific structures is less significant. The outcome does not depend much on initial conditions or on subsequent perturbations, and one can have greater confidence in one's explanation even absent data about initial conditions or other properties of the process (though of course uncertainty always remains). So stability analysis is a way to get around uncertainty.[6]

Another point worth highlighting is the high degree of idealization involved. In divide the cake, resource allocation is treated as a simple, one-shot, symmetric bargaining interaction – there are only two players, and neither has any knowledge about what the other player will demand; the network, representing social structure and connectivity, is a "boring" square lattice (there are no

[4] It is sometimes helpful to distinguish a stable state from an equilibrium. In an equilibrium the various forces in the system are, in some specifiable sense, balanced out. There are many notions of equilibrium, including well-known ones from game theory, such as the Nash equilibrium. But an equilibrium may be fragile, i.e. there are unstable equilibria. My focus here is on stability, and though stability usually implies equilibrium, I will not be concerned with this implication.

[5] One could raise doubts about whether such a tendency exists, and if so whether the game of divide the dollar represent it well. But I am interested in issues to which this matter is tangential.

[6] A caveat and a flip side. The caveat is that sometimes there are multiple stable states, in which case one needs further information to decide among them. The flipside: stability analysis provides less information about trajectories; it tells one what the end-point of a process is, but not how the process unfolded.

hubs, i.e. highly connected social nodes, for instance); the update rule is deterministic, and based on a simple kind of imitation, and so on. This kind of skeletal, simplified representation is the rule, not an exception, in this area, as in other scientific areas. There are distinct advantages to this mode of theorizing, but as we'll see later, it complicates the interpretation and the uses to which one can put the models.

Finally, to what extent is the Skyrms-Alexander model representative of work on norms in evolutionary theory? In some ways it is not representative: some models in evolutionary theory are qualitative, or not mathematical. Some do not look at individual organisms but operate at a more coarse-grained, statistical level. But the basic structure I describe above is characteristic of a broad class of models of the evolution of sociality and norms, well beyond game theoretic ones. Almost all such models depict strategic interactions (i.e. interactions in which the outcome for a player is dependent on the outcome for others). Virtually all incorporate a sketch, if not more than that, of an interaction structure; it may not be a network, or anything formal, but there is usually a specification of how social interactions come about and how members of the population are related to each other. Likewise for updating: while some models do not include an explicit specification of how each player's traits change over time, and others depict the process rather abstractly, still almost always some assumptions are made about updating. At the very least, it is implicitly assumed that updating qualitatively resembles inheritance (i.e. like begets like). Lastly, while some models explicitly depict a single organism-type, not taking into account populational variation, virtually all theoretical work in evolution accepts that it is a fundamentally population-level process.

3. Stability

So far I have discussed evolutionary models, noting that they often aim to ascertain facts about stability. But what is stability? In addressing this question, I rely on our informal understanding of the concept, but also on how the notion figures in evolutionary models like the one we just looked at. It will be seen that these notions are sufficiently close that the models can (insofar as the concept of stability is involved) be informative for normative, especially political, theorizing.

3.1. What is stability? [7] In essence, stability is a straightforward notion; it is the opposite of change. Stability (or instability) is not, however, a complete lack (or presence) of change, but rather

[7] In mathematics, specifically in so-called dynamical systems theory, there is a family of well-defined notions of stability (see Holmes & Shea-Brown 2006). Some evolutionary models explicitly rely on such concepts. While these inform my discussion here, the formal definitions are not needed for our purposes.

a graded matter: a system is more stable the less it is subject to change. Moreover, the stability of a system should not be identified with, and cannot be ascertained by, whether it *actually* changes. A system can be in a static but unstable state (i.e. in an unstable equilibrium; see footnote 4, above). So what matters is how much the system *would change if something were to impinge on it*. In other words, stability is a matter of resilience, of the potential for resisting change. Several distinctions will further clarify this notion.

The present discussion concerns the stability of social systems. In this context a distinction between two sorts of impacts or perturbations is worth drawing. There can be changes to the size or composition of a society's population – I'll call these dynamical changes. A system can be stable (or unstable) in the face of such changes – it can be *dynamically stable* (or unstable). In the example given above, this is the sort of stability at issue: does the population arrive at the same end state, in terms of its composition, no matter what the initial conditions are?

Another sort of perturbation is structural rather than dynamical. A system can be more or less stable in the face of changes to its structure. In the kinds of cases discussed here, this may involve such things as how a population is distributed in space and how members meet and interact, what kind of intrinsic factors lead it to change and so on. These are changes not to how many and which members a society has, but to how these members are organized and what they do.[8]

Stability can be further illuminated by distinguishing it from *accessibility*. Whereas stability concerns change relative to a specified state, accessibility is a matter of changes *between* states – given the state of a society at a given time, can it be "moved" to another state? If so, how likely is such a change, what costs would it entail? Etc. Now, within political philosophy, accessibility and stability are sometimes discussed together, either implicitly or explicitly, as species of the same genus – both are kinds of feasibility, i.e. they have to do with the prospects of implementing some theory or ideal.[9] In some normative respects there is indeed a close affinity between these two notions. But from the vantage point of models of the evolution of morality, stability looms larger, since, as noted above, models in this area primarily provide results about stability. So while some portions of the discussion below can be extended to feasibility in general, including accessibility, I will restrict my claims to stability and set accessibility aside.

[8] The distinction between dynamical and structural features is, to an extent, vague or pragmatic. I use it here in a way that is not affected by this.

[9] E.g. Estlund, 2011; Gilabert, 2017; Gilabert and Lawford-Smith, 2012; Lawford-Smith, 2013; Southwood, 2016. These authors discuss the notion of feasibility in terms of what is possible, or what agents (humans or institutions) 'can do'. This may explain why they pay little attention to the accessibility/stability distinction, which most readily applies to states rather than actions.

A final clarificatory comment concerns the relation between stability and homogeneity.[10] Stability, as I have characterized it, involves lack of (potential) change. Homogeneity, in contrast, consists of similarity among elements of a social state – e.g. similarity in behavioral patterns across individuals. In some cases, homogeneity leads to stability. For instance, if in divide the cake one starts with a population that consist only of demand 50%, then one stays at that state (because any "invader" than demands more or less will be worse off than the population average). However, in other cases a homogeneous population will not be stable (e.g. – a population of strictly cooperative agents in the Prisoner's Dilemma is vulnerable to invasion by defectors – Nowak, 2005 Ch. 8). The arguments I describe here appeal only to stability and may involve homogeneity, in particular cases, only to the extent that it is connected with stability (whether they do will depend on the case.) There may also be contexts where homogeneity matters for ethical purposes, but I do not discuss them here.

3.2. Stability of what? I have spoken so far of stability in a general, abstract way. But our focus here is social stability – stability of a social arrangement. What distinguishes social stability from other kinds of stability? My working assumption is that the difference lies (only) in *what* is regarded as stable – and not in what stability itself comes to.

Thus, once we have a characterization of stability, answering the question "what is social stability?", effectively requires giving a characterization of 'social', or of those aspects of the social that are relevant in the context at hand. I am not in a position to supply such a characterization – frankly, I doubt that a plausible general, context-independent characterization of 'social' is possible. In principle, one can construe the social as encompassing multiple kinds of entities (understanding 'entity' liberally) – individuals, groups and sectors, institutions, rules, norms and habits – the list can be extended and appears open ended. We can think of stability with respect to any (or many) of these entities: stability of the number and kinds of individuals, or groups; stability of the content of laws or social mores; of the structure of institutions etc. Moreover, 'the social' may have different contrasts depending on context: sometimes the social contrasts with the economic (e.g. when it is said that some group votes on 'social issues' as opposed to 'the economy'); other times the social (environment) contrasts with the natural (environment); in still other cases the focal contrast is between the social and the individual (a distinction that may partly overlap with public versus private). I do not think any one of these notions is more central or more

[10] I thank Christine Clavien for bringing this point to my attention.

appropriately thought of as genuinely social relative to others. And so I do not think there is a privileged notion of *the* social.

That said, in this paper I will focus on social stability in a fairly specific sense: the stability of a distribution of types in a population, especially behavioral types – such as the type that shares a resource or the type that cooperates. There are two reasons for this. First, since my goal is to ask whether evolutionary models can inform social and political thinking, I focus on the kind of properties which these models track. These are, as in the example above, strategic interactions like bargaining, coordination and cooperation. The second reason is that behavioral stability of this sort appears to be a necessary, or at least a very important condition, for other sorts of stability. It is hard to speak of stability of institutions or legal arrangements in a society in which basic patterns of conduct shift and change.

3.3. Stability "for the right reasons"? A third point of clarification concerns a particular notion of stability that has played a prominent role in modern political philosophy – Rawls' notion, especially as discussed in *Political Liberalism*. I have been (and will continue) speaking about stability simpliciter, whereas Rawls's discussion concerns a special notion, "stability for the right reasons" (1993, xlii). By this he means stability that is underpinned by the fact that citizens of a just society endorse basic social institutions and willingly comply with the law on moral grounds, and not due to coercion or even on the basis of strategic, real politick considerations. Perhaps the central challenge Rawls deals with in Political Liberalism is whether stability in this rich sense is possible in a heterogenous liberal society (Weithman, 2010).

Now, for Rawls this richer notion of stability contrasts with a mere *modus vivendi*, a balance of power that is not rooted in the underlying attitudes of members of society. But here my focus is on a notion that, in logical terms, is weaker than both Rawlsian stability and a modus vivendi – namely, as explained above, stability as relative lack of change. It is logically weaker because stability in this "behavioral" sense is consistent with various sorts of underlying motivations on the part of members of society. It can but need not hold "for the right reasons". I focus on this notion both because it is unclear to me whether a stronger notion of stability, specifically a Rawls-like notion, can be modeled and if so how (though see Chung, 2017). But more importantly because the arguments I will discuss shortly do not require more than a minimal, "behavioral", notion of stability. This not to say that richer notions cannot play a role in these types of arguments. It is only to say that we can and do have use for the weaker notion, and should therefore start there.

4. Arguments from stability in political philosophy

With this characterization of stability in place, I want to turn to the normative, specifically the political, significance of stability. There are two main ways, it seems, in which stability enters into debates in political philosophy. First, stability (or instability) can feature as a property of a political arrangement that can have a positive or a negative weight in assessments of it. Second, stability (or instability) can function as a constraint, a condition on a social ideal's applicability[11]. I'll discuss these in turn.

4.1. Stability as a good (or a bad). In some cases, stability plays a role as a consideration, one among many of course, for whether a proposed social arrangement is desirable. Most often, it's the price of *in*stability that is the focus. Instability impacts both the well-being of individuals and the functioning of institutions. Most fundamentally, social instability can cause material and psychological distress (though this depends, of course, on the individuals in question, as well as on the pace of change). But beyond that, instability makes planning hard, both for individuals and at the institutional level. Instability also often necessitates insurance of one form or another, a further cost. Over-stability, of course, can also be detrimental, in that it conflicts with the need for adapting to new circumstances. Whatever the case may be, we can see how stability may enter into moral-political discussions as a positive or negative feature that we may (in some cases, must) take into account when evaluating whether to accept a proposed social arrangement. In this way stability can play a role as a first-order normative consideration, a property that confers value (or disvalue) on a social-political arrangement. Or at least, it can serve as a condition affecting the value (or disvalue) of other features of a proposed social ideal: it affects the costs of implementing it. Thus, stability considerations can be an ingredient in an all-things-considered political judgment. Stability properties can make an arrangement better or worse, thus affecting its normative status.

4.2. Stability as a constraint. There is also a second, distinct, and in some ways more fundamental role for stability: it functions as a constraint. Let me introduce this with an example. A common criticism of socialism is that a socialist society is unsustainable, because, very roughly, in order to get humans to be productive members of a modern society there need to be in place the kinds of incentives that undermine the radical equality and community that socialism prescribes. For this reason, the critique goes, a truly socialist society is unfeasible. Thus socialism is not a viable social ideal and should be rejected.

[11] Here I am bracketing questions about the character of said social ideal – specifically whether it specifies a state of affairs, like a distribution of resources. Or whether it specifies principles by which to regulate an ideal society. The difference might matter for subtle issues concerning the role of feasibility and stability in political theory (Wiens, 2015), but I do not think these subtleties affects the present argument.

What interests me here is not the specific case against socialism, but the form of argument it represents. Underlying it is a political version of the maxim that ought implies can (OIC). The rationale seems to be this. A society, in particular a desirable society, is a structure that retains its basic features over time. If a social ideal, once implemented, will crumble, then by pursuing it (or at least by successfully pursuing it) we would effectively be implementing something else – a pointless exercise at best. In this kind of argument stability isn't primarily a *feature* of the arrangement under consideration, it is not something we evaluate or regard as a good or bad thing. Rather, it places limitations on the political schemes that can be carried out. In other words, in this mode we are asking whether a proposed social ideal meets a pre-condition, without which implementing it would be pointless or worse.

The constraint role of stability has a long history in political thinking. I introduced it with an example related to socialism. There are older examples. Immanuel Kant, in *Perpetual Peace,* discards the idea of a world-government because, as he puts it, its "laws [would] progressively lose their impact as the government increases its range…finally laps[ing] into anarchy" (1795\1970, p. 113). Kant is here voicing an argument of the form I described: we should not strive for a world government, he suggests, as such a government will readily fall apart, bringing us back to the imperfect, but relatively stable existing multinational order. In the background, it can be seen, is the idea that if a social-political state is inherently unstable, it cannot be desirable.

For another classic example, we can look to Hobbes. Hobbes represents an interesting case, from the present perspective. In his overall conception, we see both roles of stability combined. In very broad brushstrokes, Hobbes' view was that the only stable political regime is a strict authoritarianism – all other political arrangements would collapse into a state of nature. Therefore, such an authoritarian regime is the only one not ruled out on stability grounds. Such reasoning relies on the second, constraint role of stability. Hobbes also thought that the price of instability, a perpetual low-level war in the state of nature, is worse than living under any regime, including a coercive authoritarian one (that's the first role, in which stability is treated as a feature).[12]

4.3. A few comments. Before discussing the link between the models we looked at earlier and the arguments discussed just now, I'll offer a few clarificatory comments.

First, it is worth noting that in arguments of the sort we looked at, stability is viewed in what we may call a prospective light – we are looking ahead to a possible future state of affairs, and

[12] On some interpretations – and in some formal models that aim to capture a Hobbes' key argument – a state of nature is seen as terrible but stable. In that case stability plays only one role in Hobbes, albeit in multiple parts of his theory. As the goal is to illustrate how arguments from stability work, I think such a reading of Hobbes is less useful, though I admit that it may be less interpretively apt.

asking whether it will be stable. In contrast, I noted earlier that in evolutionary models stability is typically employed in a retrospective vein: the point of exploring whether a certain result is stable is that it provides evidence for the evolutionary model in question. This is related to a significant difference in emphasis regarding how stability enters the picture. In the second case, that of evolutionary models, what is usually sought are stability results (for epistemic reasons mentioned above). Modelers in evolutionary theory want to know which evolutionary strategies are stable and under what conditions. In the first, prospective case, *in*stability is as important, if not more important, than stability. For it is the possibility of instability that leads to worries about implementation. Nevertheless, the concept of stability is essentially the same one and this is what allows the models to inform the political argument, as I'll discuss shortly.

Second, the argument types I discuss here can, in principle, be applied at both abstract, basic principles level or at more concrete, applied levels. For instance, a stability argument against socialism operates a rather basic level, targeting a foundational conception political justice. On the other hand, in discussions of legalizing marijuana it is sometimes argued that marijuana is a "gateway drug" – that widespread use of it (following legalization) will lead to widespread use of other, more dangerous drugs. That is, that restricting drug usage to marijuana is an unstable proposal (I am not voicing an opinion on the merits of such "gateway"). Similar examples can be provided for the other kind of stability argument. The point is simply that these types argument are not inherently tied to any specific level of ethical analysis.

Third, and as noted above, my discussion of the constraint role of stability assumes a version of the ought-implies-can (OIC) maxim. While widely accepted, this maxim has been subjected to criticisms. For the most part, these depend on cases where there appears to be a "set up" such that an individual cannot fulfil a presumptive obligation – such as acting out of compulsion or Frankfurt-style cases (Vranas, 2007). Other counterexamples involve interactions between obligations, where, for instance, one person's actions are permissible as a way of preventing another person from acting impermissibly, even though the latter person may be unable to avoid performing the action in question (Graham, 2011). The arguments I look at here are not about the obligations of individuals, and it is unclear, I think, whether such counterexamples can be convincingly adapted to this context. As I do not have the space to enter into a detailed discussion of the OIC maxim, in general or in political contexts specifically, I will rely on it without further argument. Note, finally, that the other kind of argument discussed above, in which stability plays a role as a consideration for or against a social ideal, does not rely on an OIC-like principle. It is a matter of the costs and benefits of stability, to be weighed against other pros and cons.

4.4. Evolutionary models meet arguments from stability. The first main message of this paper should already be quite straightforward and apparent: evolutionary models can be relevant to political theory by providing information about stability. Such models provide us with information about the stability properties of various social arrangements – they show us whether and how stable a given state is, what kind of stability it has, which factors may lead to instability, etc. This kind of information can then be fed into arguments of the sort I charted: either ones that appeal to the price of stability and/or instability; or ones that rule out a social arrangement on instability grounds. The kind of argument to emerge, and its force, will depend on the kind of stability information provided by the model.

While this idea is, as I say, straightforward and simple given what has been said so far, it appears to have gone largely unnoticed in the existing literature on evolution and value theory.[13] Perhaps that is because of the centrality of meta-ethical issues in this literature, perhaps for other reasons. Either way it is worth noticing that evolution can link with norms in ways that do not involve moral knowledge, nor in a way that illicitly traverses the is-ought gap.

4.5. Fundamental results? Apropos the is/ought gap, I want to consider one more issue in this section, namely the possibility of a more ambitious reliance on stability, which seems to lead to stronger fundamental results in political philosophy. Primarily, I want to distinguish such a project from my own.

The issue is perhaps best approached by briefly looking at an attempt to carry out such a project, due to Ken Bimore (1994;1998; 2005). Binmore suggests to model the problem of the social contract – the mechanism by which political norms are chosen, and justified – as an extended process of bargaining. He then argues that while there are many possible solutions to such a problem, the best solution is an egalitarian one. For this, he suggests, is the only solution that is both stable (in a sense sufficiently close to the one employed here) and efficient (i.e. where resources do not go to waste). Thus, he concludes, we should adopt an egalitarian social contract, rejecting utilitarianism, Kantian morality and other systems which he views as wholly or partly at odds with a simple egalitarian conception of justice.

[13] There are a few exceptions, to be sure. A notable recent one is Kitcher (2011). Kitcher's basic view is that morality emerges—and progresses—as evolutionarily better and more stable arrangements are found, in the face of changing social challenges (especially what he calls "failures of altruism). Kitcher argues that such selected social arrangements are good in that they supply efficient solutions that preserve and enhance social productivity and stability. This is a very interesting set of ideas. But it differs from the sorts of arguments I have in mind here, as Kitcher aims to draw moral conclusions from selectionist/teleological premises. I will note that I do not find this suggestion plausible, since I think one cannot derive the appropriate sort of "ought" from teleological normativity. A fuller discussion of Kitcher's views isn't possible here.

I will not discuss the game theoretic apparatus Binmore relies on nor the specifics of his analysis. What interest me here is a deeper feature of his approach, which also marks a key difference between his use of stability and the one I discussed above: Binmore simply assumes that the task of formulating a social contrast is to be handled by applying criteria of efficiency and stability, within a rational decision-making framework, *and nothing more*. One way to see this is as the commitment of a simple is/ought fallacy. More charitably, it can be regarded as representing a staunch naturalistic stance, on which three is no more to social morality over and above the norms that evolve in what he terms The Game of Life. One way or another, in Binmore's view political morality is a set of conventions that regulate society and the distribution of resources in it, and the question facing the political philosopher is to how to find an efficient, rational (i.e. utility maximizing) and stable set of conventions. Normative questions (i.e. moral-political normativity, as distinct from the normativity of rationality) simply do not arise. In contrast, my discussion assumes that stability is *not* a fundamentally normative notion and that its role in moral-political theorizing is derived from other, more basic values.[14]

This is closely connected to another difference between the view Binmore advances and the types of arguments I have been discussing: Binmore believes one can derive positive fundamental norms (purely) from considerations of stability and efficiency – such as egalitarian distributive principles. I think that one can, at most, *rule out* basic principles—if they are found to lead to serious instability—and that, in most cases, stability issues are one among several considerations we need to take account of. More crucially, I think that stability considerations on their own cannot lead to positive normative conclusions. This is because I think of stability considerations – in the context of the OIC-based argument – as a constraint, or a kind of filter; they give us a way of narrowing down the set of options by ruling out norms that lead to instability. The candidate principles must, in my view, be grounded (at least in part) in more fundamental normative principles.

Thus, arguments I describe are on safer meta-ethical ground than Binmore's. The flip side of this safety is, of course, that the conclusions they license are less radical, according a significant but not a foundational role to stability. Stability can help us select among norms, and it can serve as a pro-tanto consideration for or against a social ideal. But it cannot, according to the present

[14] I should be clear that, in saying this, I do not rule out considerations of efficiency and utility maximization. I think that within a meta-ethical perspective that is less extreme than Binmore's they can play an important role. That role is different in some significant ways than the role of stability, however, so I will not discuss it here.

suggestion, form the very basis of our political outlook. Relative to Binmore's project, this is a modest proposal. But I think accords a far more plausible role to stability.

5. Complications (having to do with idealization)

The basic form of the arguments from stability that I've described is relatively simple. But complications arise when we look at the details, and consider how evolutionarily informed arguments from stability would actually work. The first source of complications is idealization: the tendency of evolutionary models, like models in general, to employ unrealistic simplifying assumptions.[15] The second has to do with the status of structural assumptions in these models, and whether they have a normative dimension. In this section and the next I describe these issues in turn and make some suggestions about how they may be handled.

The Skyrms-Alexander example is rife with idealizations, as I noted when introducing it. Among them: the distribution problem (modeled by the game of divide the dollar) is exceedingly simple – it involves dividing a unit good among two symmetrically situated agents. Social structure is represented by a regular lattice, i.e. every individual has the same number of social connections, and these have the same weight in the individual's interactions. And social learning works solely by success-driven imitation, with information restricted to the learner's immediate environment. These simplifications may make it unclear which situations in the world a given model applies to. Even when applicability is clear, there are usually important mismatches between the model and the world, and these make it hard to assess the value of the model's predictions. One might legitimately worry that this relegates stability results to an idle, merely in-principle role: *If* we could learn about real-world stability from evolutionary models that might affect our normative thinking. But we cannot learn much, because evolutionary models are too far removed from reality. These kinds of concerns arise in many areas, and I think they should be taken very seriously. I will discuss two strategies for dealing with them. I do so under the assumption that a straightforward filling-in or "de-idealization" of the models is not possible – we cannot, in most cases, revert to a model that does not contain idealizations.

[15] 'Idealization' is naturally reminiscent of the term 'ideal', in the normative sense (as in "the ideal of equality"). But the terms pick out different notions: 'idealization' refers to the scientific practice of introducing known distortions into a model (Levy, 2018). An 'ideal' is a principle that should regulate conduct. That said, there are interesting connections between idealization and theorizing about ideals, e.g. inasmuch as normative theorizing about ideals engages in idealization (Enoch, 2005), or in debates surrounding so-called 'ideal theory' in political philosophy (Valentini, 2012). Engaging with these issues will take me beyond the scope of the present paper.

The first strategy involves treating the model as what some philosophers of science call a *minimal model* (Batterman & Rice, 2014; Weisberg, 2007). A minimal model isolates a capacity or a causal tendency, often making radical simplifications for this purpose. Thus, consider Thomas Schelling's famous model of racial segregation (Schelling, 1971). In this model, there are two kinds of agents, and they are distributed across a grid representing an urban residential environment. Each agent prefers to reside in a neighborhood (an area on the grid) in which members of his group are present in at least some fraction *f*. If the fraction of like neighbors falls below *f*, the agent moves to a neighborhood where this condition is satisfied. Schelling showed that in such a model segregation can arise even if *f* is not very high (about ⅓). This is seen by many as a minimal model for bottom-up segregation – segregation that results from the actions of individuals, and not from central planning of any sort. And, moreover, it can arise even when none of the individuals have a preference for living in a segregated city. But the model clearly makes many idealizations, and cannot be seen as anything like an accurate representation of a real-world urban environment.

Now, when we come to understand a real-world context, minimal models are not applied as is. Instead they tend to play a guiding role. The minimal model informs us about the kinds of factors at play and the in-principle tendencies they have – the Schelling model shows how a process of "segregation nobody wants" can occur, given a certain setup. Once an abstract simple scenario like the Schelling model is well-understood, real-world consequences are explored via context-specific models, which make assumptions appropriate for the particular case at hand, such as specific residential patterns, more complex preferences, etc. Often, this requires empirical information, but also analytical tools for studying a more complex model. Thus, a qualitative tendency is gleaned from the minimal model, and then a context-specific prediction is made by adapting it to the relevant context (such a prediction can then be used for confirmation or other purposes).

Skyrms' model, to return to the example we began with, shows us that an environment where bargaining-like interactions are primarily local – i.e., restricted to an agent's close neighborhood – and in which social learning has a high-fidelity payoff-based character, tends to stabilize equal sharing. This is a non-trivial result. But it is not applicable as is, since few if any real-world situations involve pairwise bargaining in a simplified lattice-like structure that meets the strictures of divide-the-dollar. It is possible, however, to seek contexts with these general features – a highly connected social network where interactions are primarily among neighbors, distributive problems with small numbers of participants, such that equal sharing can be efficient, and so on – and look at the stability of equality within them. This way we get closer to a real-world "verdict" about the stability of egalitarian arrangements.

A second strategy to deal with idealization is to compare different models. This is sometimes referred to as *robustness analysis*, since it's meant to test whether the results of the model are robust to various perturbations (Weisberg, 2006; Wimsatt, 1981). Robustness analysis is usually most relevant for testing structural assumptions. Suppose a class of models vary in some structural respect, such as the type of network employed or the character of the update rule. Nevertheless, these models provide similar results. If the variation in the models covers enough of the space of possible structures, this can be seen as an argument that such variation does not make a significant difference, and that the result is independent of the specific assumptions made in one or another of the models. In other words, suppose one can show that in a Skyrms-like setup, a diverse set of assumptions about network structure all yield the result that an equal sharing equilibrium is stable. Then one can be fairly confident that the network structure of any particular model, even if highly idealized, does not matter a great deal.

It is often difficult to demonstrate robustness. Sometimes, it is even difficult to formulate the models needed for comparison. Considerable simulation work may be needed, to test a sufficiently wide range of structural assumptions. But if the right range of models is compared, and if significant results can be obtained, then this method allows one to get from a family of models, each of which is partial and idealized, to a conclusion that is, in a sense, independent of the idealizations and limitations of each model on its own.

The Skyrms model has indeed undergone some testing along these lines. Work contained in Alexander (2007), suggests that the precise details of the learning dynamics do not matter: equal splitting will arise so long as learning is some sort of payoff-based imitation. However, this same work demonstrates the flip-side of robustness analysis. For Alexander also shows that some types of learning will *not* lead to equal splitting. This holds primarily for learning rules that are not imitation based: for instance, best-response learning (in which the agent makes a prediction as to the behavior of its partner, then chooses the strategy that will maximize payoff relative to that prediction), can lead *away* from an equal split. Such a result is also useful, of course, though it demonstrates reduced robustness. So robustness analysis can buttress, but also undermine, a stability result.

Now, the issue of idealization is important but it is also, as noted, general: idealized models are found across a wide range of areas, and the methods I have described are common strategies for addressing them. These issues have less to do with the application of evolutionary models to normative matters and more to do with the fact that they are idealized models. The next section looks at an issue that is specifically relevant to the role these models are meant to play in a political-normative context, and whether or not they can play the role I've ascribed to them.

6. Further complications (having to do with structural assumptions)

Any interesting evolutionary model will make substantial structural assumptions. We saw as much in our discussion of the Skyrms-Alexander example. That model made assumptions about how the social network is connected: it is regular, i.e. connections are distributed homogenously, rather than in a hierarchical or center-periphery manner, for instance. And it portrays social learning in a specific way, i.e. as a form of imitation, driven by success. In modeling, making some such assumptions is inevitable. And, often, the richer and more specific the assumptions, the greater the power of the model to provide definite, interesting results.

Ordinarily, you will recall, evolutionary models get used in a descriptive-retrospective mode. In this context, such assumptions have a purely descriptive character. They aim to capture the way the world is, or was. This is often done in an approximate and idealized way but the direction of fit, as it were, is model-to-world. However, in the kinds of arguments I have been discussing, models are treated prospectively; they are put to use in a forward-looking normative context. In such a case, structural assumptions are part of representing an ideal, a putatively desirable social state. The direction of fit, then, is world-to-model. But this raises the possibility that structural assumptions that are acceptable from a descriptive standpoint, inasmuch as they are an appropriate depiction of how things actually stand, are objectionable from a normative standpoint. To put the point differently, a stability-based analysis has a conditional form. It says, in effect: if a society has such and such a structure, then it will be stable in the face of such and such perturbations. Put to normative use, the antecedent is taken to apply to (a proposal for) an ideal society. But then one may reject the argument by rejecting the antecedent *on normative grounds*. Thus, suppose someone argues for rejecting a social arrangement due to its instability. And suppose, further, that the instability result she relies on is derived from a model in which updating occurs via imitation. In this kind of case one could, in principle, object to the instability argument by countering that in an ideal society citizens would be much less inclined to imitation learning (perhaps because they will be better educated, better informed through their own cognitive efforts, more critical of peers and so on). Similarly, different forms of social connectivity may be objectionable to some. One might favor a society in which people break out of their close environments, for instance, or in which there is less homogeneity of social structure. Someone holding such view would dismiss arguments from stability whose basis is, in part, assumptions about social structure that confine agents to a local environment. In other words, modeling assumptions may have normative aspects, such that claims that look perfectly reasonable from a descriptive standpoint, adequately depicting the reality of the social systems in question, may be called into question when applied to normative issues.

This is a general version of a kind of issue – a kind of counter-argument to stability concerns – that has not gone entirely unnoticed. For instance, in his posthumously published essay "Why not Socialism?" G.A. Cohen argues as follows. It is often suggested, he notes, that the main obstacle to implementing socialism is its clash with basic features of human desires and motivations. But Cohen thinks this begs the question against the socialist. The real problem, he thinks, is that we do not know which "social technologies" would enable a socialist society to succeed and persevere (Cohen, 2009, Ch. IV). By 'social technologies' Cohen appears to mean a broad range of institutional and organizational features of society, including features of the sort I have labeled 'structural' in the evolutionary models we have discussed.

The idea underlying Cohen's argument, it appears, is that critiques alleging that socialist societies are unstable presuppose an ethos and an incentive structure that are characteristic of capitalist societies. But, Cohen asks, why should a socialist accept this assumption? The socialist rejects capitalism, including (perhaps first and foremost) its ethos and its view of human psychology.[16] Of course, as he acknowledges, the point is symmetric. The socialist needs to offer a credible account of the social technologies needed to sustain socialism. Thus, to put Cohen's point slightly differently, both in order to make a (compelling) stability argument and in order to reject such an argument, one needs to say something about the "social technologies" required for a stable socialist society.

The broader point isn't specific to socialism, of course. It is that when one brings descriptive knowledge (about stability) to bear on normative questions, social technologies and other structural-descriptive features may be subject to normative evaluation. Assumptions about structure should not be taken to enjoy "descriptive immunity" – they may well be subject to normative evaluation, and such an evaluation may well lead us to reject or revise the argument.

What is the impact of this issue on the utility and relevance of arguments from stability? Several points can be made. First, if structural assumptions are disputed (and assuming such assumptions are required for the relevant modeling results) then the scope of the argument will be restricted – it will be accepted only by those who accept the structural assumptions. This need not result in a preaching-to-the-choir situation, since it is possible that holders of overall quite different political philosophies will nevertheless agree on the desirability (or otherwise) of certain structural features. For instance, both a democratic-socialist and free-market enthusiast can agree that individuals should be independently minded and avoid blindly imitating others in society. So both

[16] "The easiest way to generate productivity in a modern society is by nourishing the motives… of greed and fear [which] are repugnant motives." (Cohen, 2009, 75-6).

can accept a certain picture of social learning, such that even if certain modeling results depend on such a picture, this will not pose a problem.

A second point is that, quite obviously, an argument from stability can be supplemented with an argument for the structural assumptions that underlie it. Perhaps one can make a case for social connectivity being homogenous sans any commitment to disputed principles. If so, one can use this to buttress an argument from stability, securing its modeling assumptions.

A third and final point is that the question of structural assumptions may interact with an issue discussed earlier, namely robustness analysis. If one can show that disputed structural assumptions do not matter for the (in)stability result—if one can demonstrate structural robustness with respect to them—then one need not worry whether they are acceptable or not, and the argument from stability can go through regardless. But of course, the proof of such a pudding is in the eating: one would need to perform the analysis and demonstrate robustness.

Having made these remarks, I do not wish to suggest that the status of structural assumptions is simple or unimportant. To the extent that such assumptions are disputed, so will the associated stability argument. My intention is not, by any means, to suggest that use of evolutionary models is a silver bullet in assessing social stability, nor that such use is simple or problem free. My goal in this paper is to flesh out stability-based arguments and to discuss how they work. The issue of structural assumptions represents a (significant) wrinkle in applying such arguments, and making such wrinkles salient is part of that goal.

7. Conclusion

Let me offer a brief summary. I have argued that there is a coherent in-principle path from stability results in evolutionary modelling to moral-political conclusions. More specifically there are two related paths: one in which stability enters as a first-order consideration, having to do with the value or disvalue of stability; and one in which it serves as a constraint on the applicability of political schemas, grounded in a version of the ought-implies-can principle (both rely on stability, albeit in distinct ways.) Such argument patterns do not involve an (illegitimate) move from is to ought, and they do have bite – i.e. they have the potential to provide us with interesting critical information.

I've also tried to assess the applicability of the in-principle argument, looking at some potential complications. One class of complication concerns idealization. It might be dealt with either by treating the models as minimal, or by way of robustness analysis. Another class stems from the fact that structural assumptions can be viewed as normatively questionable. I am less confident about whether and how this can be fully circumvented, short of studying models that make un-

objectionable assumptions. But I offered some remarks about how models can be relied on in spite of disputed structural assumptions. Bearing in mind the complications that I've detailed may affect not only how one uses existing models, but also which models the field goes on to develop.

References

Alexander, J.M. (2007). *The Structural Evolution of Morality*. Cambridge University Press.

Batterman, R. & Rice, C. (2014). Minimal Model Explanations. *Philosophy of Science* 81(3): 349-76.

Binmore, K. (1994). *Playing Fair: Game Theory and the Social Contract I*. MIT Press.

Binmore, K. (1998). *Just Playing: Game Theory and the Social Contract II*. MIT Press.

Binmore, K. (2005). *Natural Justice*. Oxford University Press.

Chung, H. (2017). The Instability of John Rawls's "Stability for the Right Reasons". *Episteme*, online at: https://doi.org/10.1017/epi.2017.14

Enoch, D. (2005). Why Idealize?. *Ethics* 115(4), 759-787.

Enoch, D. (2010). The Epistemological Challenge to Metanormative Realism: How Best to Understand it, and How to Cope with it. *Philosophical Studies* 148 (3):413-438.

Estlund, D. (2011). Human Nature and the Limits (if Any) of Political Philosophy. *Philosophy & Public Affairs*, 39(3): 207-237.

FitzPatrick, W. (2014). "Morality and Evolutionary Biology", *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/morality-biology/>.

Gilabert, P. (2017). Justice and Feasibility: A Dynamical Approach. In K. Vallier & M. Weber (eds.), *Political Utopias: Contemporary Debates*. Oxford University Press.

Gilabert, P. & Lawford-Smith, H. (2012). Political Feasibility: A Conceptual Exploration. *Political Studies* 60 (4):809-825.

Graham, P.A. (2011). 'Ought' and Ability. *Philosophical Review*, 120(3): 337-382.

Harms, W. & Skyrms, B. (2008). Evolution of Moral Norms, in Michael Ruse, ed., *The Oxford Handbook of Philosophy of Biology*. Oxford University Press.

Holmes, P. & Shea-Brown, E.T. (2006). Stability. *Scholarpedia*, 1(10):1838.

Joyce, R. (2006). *The Evolution of Morality*. MIT Press.

Kitcher, P. (2011), *The Ethical Project*. Cambridge, MA: Harvard University Press.

Lawford-Smith, H. (2013). Understanding Political Feasibility. *The Journal of Political Philosophy*, 21(3): 243–259.

Levy, A. (2011). Game Theory, Indirect Modeling and the Origins of Morality. *Journal of Philosophy*, CVII (4):171-187.

Levy, A. (2018). Idealization and Abstraction: Refining the Distinction. *Synthese*, https://link.springer.com/article/10.1007%2Fs11229-018-1721-z.

Lewens, T. (2015). *Cultural Evolution*. Oxford University Press.

Ruse, M. (1986). *Taking Darwin Seriously*. Prometheus Books

Ruse, M. & Wilson E.O. (1986). Moral philosophy as Applied Science. *Philosophy* 61: 173-192.

Skyrms, B. (1996). *Evolution and The Social Contract*. Cambridge University Press.

Skyrms, B. (2003). *The Stag and the Evolution of Social Structure*. Cambridge University Press.

Southwood, N. (2016). Does 'Ought' Imply 'Feasible'? *Philosophy & Public Affairs* 44(1): 7-45.

Street, S. (2009). Evolution and the Normativity of Epistemic Reasons. *Canadian Journal of Philosophy* 39: 213-248.

Valentini, L. (2012), Ideal vs. Non-ideal Theory: A Conceptual Map. *Philosophy Compass* 7(9): 654-664.

Vranas, P.B.M. (2007). I Ought, Therefore I Can, *Philosophical Studies*, 136(2): 167-216.

Wiens, D. (2015). Political Ideals and the Feasibility Frontier, *Economics and Philosophy* 31(3): 447–477.

Weisberg, M. (2006). Robustness Analysis. *Philosophy of Science* 73(5): 730-742.

Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy* 104(12): 639–659.

Weithman, P. J. (2010). *Why Political Liberalism? On John Rawls' Political Turn*. Oxford University Press.

Weibull, Jörgen W. (1995), *Evolutionary Game Theory*. Cambridge, MA: MIT Press.

Wimsatt, W. (1981), Robustness, Reliability and Overdetermination," in M. Brewer and B. Collins (Eds.), *Scientific Inquiry and the Social Sciences* , (San Francisco: Jossey-Bass).

Young, H.P. (2004). *Strategic Learning and its Limits*. Oxford University Press.