Seven Problems with Massive Simulation Models for Policy Decision-Making

Till Grüne-Yanoff [1,*]

Email gryne@kth.se

[1] KTH Stockholm, Royal Institute of Technology, Brinellvägen 32, 10044 Stockholm, Sweden

Abstract

Policymakers increasingly draw on scientific methods, including simulation modeling, to justify their decisions. For these purposes, scientist and policymakers face an extensive choice of modeling strategies. This paper distinguishes two types of strategies: *Massive Simulation Models* (MSMs) and *Abstract Simulation Models* (ASMs), and discusses how to justify strategy choice with reference to the core characteristics of the respective strategies. In particular, I argue that MSMs might have more severe problems than ASMs in determining the accuracy of the model; that MSMs might have more severe problems than ASMs in dealing with inevitable uncertainty; and that MSMs might have more severe problems than ASMs with misinterpretation and misapplication due to their format. While this in no way excludes the prospect that some MSMs provide good justifications for policy decisions, my arguments caution against a general preference for MSM over ASMs for policy decision purposes.

1. Introduction

Policymakers and their advisors today face a bewildering array of models to choose from. Most obvious, perhaps, is their choice between different structural and causal features, as well as between different parameterizations for each. On another level, they also need to choose how much detail their models shall incorporate. Recent technological advances have rapidly expanded the amount of detail that can be processed when computing a model. Models whose detail is constrained largely by current computational capacities I call *Massive Simulation Models* (MSMs). Models whose detail is also constrained by other considerations (of simplicity, of transparency, etc.) I call *Abstract Simulation Models* (ASMs). Although simulation models can, of course, be distinguished further, this simple dichotomous distinction will suffice for the present argument.

It might seem an obvious and trivial claim that policymakers should prefer MSMs over ASMs for most if not all purposes: MSMs, because they contain more detail, can be closer approximations to the real system; they can better represent complexity and population heterogeneity; and the policymaker can use it as a holistic test-bed for potential policy interventions. ASMs, due to their additional constraints on detail, do not offer the same potential as MSMs in these regards—and therefore might be regarded as inferior for policy purposes.

Contrary to that claim, I will argue in this paper that for many policy purposes, ASMs are preferable to MSMs. Essentially, my argument is that, although MSMs have larger potentials than ASMs in various dimensions, they are also more likely to fail—and that in many cases, this probability of failing outweighs their higher potential. Because this argument is difficult to make in

full generality, I will focus on one case here, that of smallpox vaccination policies. By comparing one specific MSM and one specific ASM, I show that for some policy purposes, the ASM is preferable to the MSM.

The paper is structured as follows: Sect. 2 gives a conceptual distinction between MSMs and ASMs; Sect. 3 illustrates this difference with a case from the smallpox vaccination literature; Sect. 4 discusses the seven problems with MSMs that make ASMs preferable for some policy purposes; and Sect. 5 concludes.

2. MSM vs. ASM

In this paper, I distinguish *Massive Simulation Models* (MSMs) from *Abstract Simulation Models* (ASMs). By simulation model, I mean any dynamic model that represents a target and that is solved through some temporal process (Humphreys 2004, p. 210). For most simulation models, these solution processes also have a representational function—they are interpreted as a representation of the dynamic processes or mechanisms operating in the target (Hartmann 1996, p. 83). Many such models, and in particular MSMs, are implemented on a computer, but this is not an essential property. Instead, some ASMs are also realized materially (like Schelling's checkerboard model, Schelling 1971) or as paper-and-pencil models.

MSMs differ from ASMs at first glance by their much higher level of detail, especially the number of variables and parameters they include, and the number of relation between these. In the case that I discuss below, for example, the MSM includes approximately 1.6 million vertices with maximally 1.5 million edges that might change 24 times a day; while the smaller ASM includes approximately 7900 vertices with maximally 6000 edges (and approximately 55,000 vertices for the larger one) that can change maximally 17 times a day.

Based on the richness in realistic detail, MSMs are often claimed to offer a highly accurate picture of the real system:

> Such models allow for the creating of a kind of virtual universe, in which many players can act in complex—and realistic—ways. (Farmer and Foley 2009, p. 686)

Understood in this way, MSMs are typically seen as direct representation of real systems: their structure allows for a mapping from the model to the target without having to take recourse to mediating models. ASMs, in contrast, can hardly ever claim to represent a real system directly—their level of detail is not sufficient. At best, they are able to represent *stylized facts* about or abstractions of a system, which have been prepared through an abstraction or idealization procedure from the real system.

That MSMs represent real systems directly is further supported by the practice of fitting or calibrating the model directly to real data. The MSM I discuss below, for example, bases the specification of edges and their hourly change on census data. The authors of the comparison ASM, in contrast, stipulate certain edge changes and interpret these activities as "being at home," "going to work," "being at the hospital," etc. In the case I discuss here, these interpretations are not based on data from actual target systems, but rather rely on plausibility intuitions. Alternatively, they could have been based on interceding models, which abstract stylized facts of "going to work" or "being at the hospital" from available data, and then allow the representation of these

abstractions in the ASM.

Both MSMs and ASMs typically represent processes of their target systems. But even here, there is an important difference. MSMs typically represent a multitude of simultaneous processes or mechanisms, while ASMs typically represent only one or a small number of such processes. Take, for example, the following claim about the advantages of agent-based models in economics:

A thorough attempt to understand the whole economy through agent-based modeling will require integrating models of financial interactions with those of industrial production, real estate, government spending, taxes, business investment, and with consumer behavior. (Farmer and Foley 2009 , p. 686)

Presumably, many of these components will operate through different mechanisms. Consequently, putting them all together in a single MSM implies that many different processes will operate simultaneously when producing a model outcome. ASMs, in contrast, will typically focus on a small subset, or even just one, of these processes.

So far, I have distinguished MSMs and ASMs only with respect to their different representational relations to a target. Now I also wish to distinguish them with respect to how these relations are interpreted and for what purposes they are used. Regarding interpretation, the greater amount of detail in their models is typically employed by MSM modelers in order to achieve a "realistic" interpretation of the model's representational function. "Realisticness" is a subjective psychological effect that might stem from an impression of familiarity and might lead to greater trust in the model and its conclusions:

[D]ecision makers might be more willing to trust findings based on rather detailed simulation models where they see a lot of economic structure they are familiar with than in general insights obtained in rather abstract mathematical models. (Dawid and Fagiolo 2008 , p. 354)

ASMs do not offer such a rich collection of familiar details, are therefore typically not considered as "realistic" as MSMs, and thus might not inspire the same amount of trust.

The differences discussed so far also imply an important difference in the use of these model types for policy decision-making. MSMs are often used as part of a "holistic approach": a "model of the whole economy" (Farmer and Foley 2009 , p. 686) is used as a "virtual universe" (ibid.) to evaluate the effects of proposed interventions in the target system. That is, interventions are simulated in the model, and model results are interpreted as forecasts of the results of such interventions in the real system. ASMs cannot be used in this way, as they do not offer a representation of the whole system or of the combination of its many operating mechanisms. The holistic synthesis that the MSM promises as part of its package must be performed by the ASM user in some other way, for example through expert judgment.

To be clear, MSMs and ASMs have important similarities, despite the discussed differences between them. Both aim to represent non-linear and complex behavior, albeit on different levels of abstraction and idealization. Furthermore, both types of models abstract and idealize, but to different degrees, and for different reasons. MSMs abstract and idealize for tractability and computation reasons: they are mainly constrained by current computational capacities. ASMs, in

contrast, are also constrained by other considerations (for example, of simplicity, of transparency, etc.), so that computability rarely becomes a relevant constraint for them.

Last, the distinction between MSM and ASM is itself a simplification. Many actual simulation models exhibit some properties of the one kind and some of the other, and thus do not clearly fall into either category. However, this does not pose a problem for my argument; my discussion in Sect. 4 addresses these properties separately so that relevant conclusions can be drawn for such "in-between models," too. While I am aware of the possibility of such cases, I have nevertheless decided to stick with the dichotomous distinction for ease of exposition.

3. The Case of Vaccination Policy Modeling

Vaccination is one of the most effective ways of fighting epidemics. However, many vaccines do not provide long-term protection, or have serious side effects, so that a preventive vaccination (e.g., to all children aged five) is not feasible. Instead, these vaccines should be applied only when the risk of an epidemic is sufficiently high. The policymaker then has to make a momentous decision: namely, how to apply vaccinations in a large population when an epidemic is imminent or has already broken out. The most relevant alternatives are a *tracing vaccination* (TV), where the potential recent contacts of an infected individual are traced and vaccinated; a *limited vaccination* (LV), where a random subset of the population is vaccinated; or a *mass vaccination* (MV), where the whole population is vaccinated. The choice is not trivial: MV is more likely to stop the spread of the disease, but is more costly and bears vaccination risks for a large population, while the effectiveness of LV and TV is less certain, but they are less costly, and do not expose a large number of people to vaccination risk. An instructive discussion in this regard is the evaluation of vaccine policies in the UK during the 2009 influenza pandemic, when an LV was implemented (Hine 2010).

AQ1

Various modeling projects have sought to contribute to this policy decision (for an overview, see Grüne-Yanoff 2011). Early attempts tried to model the epidemic dynamics as an aggregate equation. For example, Kaplan et al. (2002) simulate an attack of 1000 initial smallpox cases on a population of 10 million. The population is assumed to mix homogeneously—i.e., to consist of identical individuals who have an equal chance of interacting with any other member of the population. R_0, the rate of infections a single infectious agent generates among susceptibles, is assumed to be uniform throughout the simulation. $R_0 = 3$ is derived from historical data. An infected agent undergoes four stages. Only in the first is she vaccine-sensitive; only in the third and fourth is she infectious; in the fourth, however, she shows symptoms (scabs) and is automatically isolated. Additionally, the administration of vaccinations is modeled under logistical constraints: MV of the whole population is achieved in 10 days. Tracking *and* vaccinating an infected person in TV, however, takes four times as many nurse-hours as a simple vaccination.

Kaplan et al. (2002) thus offer an example of an equation-based simulation study. By assuming homogeneous mixing, the infection rate R_0 becomes a parameter characterizing the population. Policy effects are then modeled directly on this population parameter, and the main question is whether vaccine administration can outpace the random spread in the population. Unsurprisingly, perhaps, the results heavily favor MV over TV. Initiated on day 5 after the initial attack, MV leads to 560 deaths, while TV leads to 110,000 deaths. Sensitivity analysis shows that TV is more sensitive than MV to the size of initial attack and changes in R_0, further supporting the strong results in favor of MV. The time to identify and then vaccinate the exposed is simply too long for the specified R_0 and for the assumed period in which the exposed are still sensitive to the vaccine.

Models like Kaplan et al. (2002) have been criticized for their homogeneity assumption. For smallpox infection, close and extended contact between infected and healthy agent is required. In a population of 10 million, it is highly implausible that an infected agent has the same probability of having contact with any non-infected members. Furthermore, how the infected agents move through the population—i.e., with whom and with how many healthy agents they have contact—might influence the effects of different vaccination policies. To model this contact pattern was the main motivation behind the developments of various agent-based models. I will discuss two instances here, one ASM and one MSM.

My example of an ASM, Burke et al. (2006), simulates a single initial infected person attack on a town network of either 6000 or 50,000 people. Town networks either consist of one town (uniform), a ring of six towns, or a 'hub' with four 'spokes.' Each town consists of households of up to seven persons, one workplace, and one school. All towns share a hospital. Each space is represented as a grid, so that each cell in the grid has eight neighbors. Agents are distinguished by type (child, health care worker [5% of adult population], commuter [10%], and non-commuter [90%]) by family ID and by infectious status. Each 'day,' agents visit spaces according to their type, and then return home. On the first 'day' of the simulation, the position in schools and workplaces is randomly assigned, but after that, agents remember their positions. During the 'day,' agents interact with all of their immediate neighbors: 10 times at home, 7 times at work, and 15 times in the hospital. After each interaction, they move positions to the first free cell in their neighborhood. Homogeneous mixing is thus completely eschewed; instead, agents interact in a number of dynamic neighborhoods.

Transmission occurs at a certain rate in each of the agents' interactions. It can infect both contactor and contacted. Transmission rates depend on the stage the infectious person is in, the type of disease he has, and whether the susceptible agent has partial immunity.

Burke et al. assessed only TV as a first policy intervention, and LVs of varying degrees only as 'add-on' measures. Results for all three town networks showed substantial concordance. Contrasted with a 'no response' scenario, TV in combination with hospital isolation was sufficient to limit the epidemic to a mean of fewer than 48 cases and a mean duration of less than 77 days. Post-release LV of either 40% or 80% of the total population added some additional protection, reducing the mean of infected people to 33 and shortening the mean duration to less than 60 days.

My example of an MSM, Eubank et al. (2004), simulates an attack of 1000 infected agents on the population of Portland, OR, of 1.5 million. Portland is represented by approximately 181,000 locations, each associated with a specific activity, like work, shopping, or school, as well as maximal occupancies. Each agent is characterized by a list of the entrance and exit times into and from a location for all locations that person visited during the day. This huge database was developed by the traffic simulation tool TRANSIMS, which in turn is based on US census data.

Smallpox is modeled by a single parameter, disease 'load' (analogous to a viral titer). Agents have individual thresholds, above which their load leads to infection (and load growth at individual growth rates), symptoms, infectiousness, and death. Every hour, infectious agents shed a fixed fraction of their load to the local environment. Locations thus get contaminated with load, which is distributed equally among those present. Shedding and absorption fractions differ individually. Infected individuals withdraw to their homes 24 h after becoming infectious.

The Eubank et al. (2004) model is deterministic. MV with a 4-day delay resulted in 0.39 deaths

per initially infected person; TV with the same delay in 0.54 deaths. Varying delays, they found that delay in response is the most important factor in limiting deaths, yielding similar results for TV and MV.

Both papers give more or less unconditional policy advice. To quote just two examples: "[C]ontact tracing and vaccination of household, workplace and school contacts, along with effective isolation of diagnosed cases, can control epidemics of smallpox" (Burke et al. 2006, p. 1148); and "[O]utbreaks can be contained by a strategy of targeted vaccination combined with early detection without resorting to mass vaccination of a population" (Eubank et al. 2004, p. 180).

I classify Eubank et al. (2004) as a MSM and Burke et al. (2006) as an ASM. The former incorporates much more detail than the latter and is widely considered to be more realistic. The former is also proclaimed as a direct representation of a real system, namely the city of Portland, and it is based on and calibrated with census data from that target system. The latter is not claimed to represent any concrete system, nor does it make use of any data. Instead, it explicitly claims to represent an "artificial city" that shares some properties with real cities, but is different otherwise (Burke et al. 2006, p. 1142). Furthermore, the former includes many more simultaneous mechanisms than the latter: it distinguishes several activities at each location, each of which yields different contact rates; it also includes the effects of demographic factors (age in particular) on mixing; it distinguishes different forms of smallpox; and it tries to incorporate at least some rudimentary effects of infection on behavior. The latter model included a lesser number of locations and did not distinguish activities or demographics, nor did it include infection effects on behavior. Finally, while the Eubank et al. (2004) model at least implies a holistic approach, the Burke et al. (2006) has no such aspirations. One could therefore conclude, *prima facie*, that the former is a more powerful tool for deciding vaccine policies than the latter. In the following section, I will argue that this is not necessarily so.

4. Seven Problems with MSMs for Policy Purposes

This section discusses seven problems that show why, in some situations, an ASM might be preferable to an MSM for policy purposes. These problems are conceptually separate, although in practice they often overlap.

4.1. What Is the Target?

Prima facie, MSMs like Eubank et al. (2004) have a particular target: for example, the town of Portland, OR. ASMs like Burke et al. (2006), in contrast, do not appear to have such a particular target; rather, they represent an abstracted type, like "a town" or "an urban population network." Consequently, MSMs are often judged to be more *realistic* than ASMs, as model users can more easily trace the MSM features to the properties of a particular target. This realisticness judgment, in turn, as the above quote from Dawid and Fagiolo (2008) shows, often induces policymakers to place more trust in the reliability and usefulness of the model in question. For this reason, MSMs often seem preferable to ASMs for policy purposes.

But is the inference from realisticness to reliability and usefulness justified? Presumably, the argument is that (1) judging a model to be realistic indicates that it is a highly accurate representation of the target, and that (2) highly accurate representation of the target is a necessary condition for the model to give reliable and useful information about possible policy interventions in the target.

While I do not dispute these claims individually here, I argue that their conjunction does not constitute a valid argument if the meaning of "target" changes between them. This is precisely what happens in the smallpox simulation studies. The target *of the policy question* is the city environment generally, as the introductory sentence of Eubank et al. (2004 , p. 180) shows:

The dense social-contact networks characteristic of urban areas form a perfect fabric for fast, uncontrolled disease propagation. […] How can an outbreak be contained before it becomes an epidemic, and what disease surveillance strategies should be implemented?

Furthermore, because epidemic policies are typically the responsibility of national or international institutions, the targets of the policy question are all cities within the governing domain of that institution (e.g., all US cities, all cities in industrialized countries, all cities of the world, etc.). The target of such a policy question thus is an abstract entity: the network characteristics of all urban areas within the relevant domain.

The *model's target* in the MSM case, in contrast, is a particular: the city of Portland, OR. The authors of this model suggest that it is just an instance of the network characteristics in urban areas.[1] But by choosing a particular target, they allow for a possible divergence between the meaning of "target" in step (1) and (2) in the above argument. In particular, the judgment that their model is realistic might now be based on relational features of their model and the city of Portland that are wholly irrelevant for the relational features of their model and network characteristics of all urban areas within the relevant domain. For example, inclusion of the Columbia riverbed, of the locations of Portland's universities, as well as Portland's public transport system, might increase the realisticness of the model. However, these might be features that are either *irrelevant* for the path of an epidemic through an urban network, or they might not be representative of urban networks in the US more generally. Both of these cases might sever the relation between realisticness, reliability, and usefulness: an MSM with these features might be more realistic than a ASM, while the ASM is more accurate representation of the general network characteristics of all urban areas within the relevant domain. In such cases, the ASM would be a more powerful policy tool than the MSM.

4.2. How to Measure Parameters

MSMs differ from ASMs in their much higher level of detail, especially the number of variables and parameters they include, and the number of relation between these. Assuming that both models have the same target (so that problem 4.1 does not arise), a higher number of variables and parameters gives MSMs more potential than ASMs to accurately represent the target system. *Prima facie*, this gives MSMs an advantage over ASMs for policy purposes.

However, this argument assumes that the additional variables and parameters that give MSMs an advantage over ASMs can be measured or estimated with sufficient accuracy. Both of these assumptions are problematic. I will discuss measurement problems in this subsection and estimation problems in the next.

The measured variables and parameters of the smallpox MSM are those whose value is directly obtained from some external data source. For example, properties like age, occupation, health, and home location are obtained from census data for all of the 1.5 million agents in the model. Properties of the urban transport network and of land occupation and use are obtained from urban

planning organizations (Eubank et al. 2004 , Supplement, 3). These examples of massive data intake seem indeed to support the comparative detail richness of MSMs over ASMs.

However, a closer reading of the article and its supplementary material reveals that many of the parameters and variables could not be accurately measured (or even measured at all). Instead, they are determined by *ad hoc* assumptions, best guesses, or the use of reasonable ranges. I describe three instances here for illustrative purposes. The first concerns the disease-relevant contacts of agents within a location, which cannot be found in census data:

We do not have data for proximity of people, other than that they are in the same (possibly very large) location. […] It seems as though the dependence on distance is very coarse: one mode of transmission occurs at close ranges (< 6 feet) and another for large ranges. We have developed an *ad hoc* model that takes advantage of this coarseness. (Eubank et al. 2004 , Supplement, 9)

This *ad hoc* model makes uniform assumptions about the occupancy rate of locations within a city block that are, the authors admit, "nothing more than reasonable guesses" (Eubank et al. 2004 , Supplement, 11). Location occupancy rates, however, crucially influence the number of possible contacts—and hence may be relevant for the spread of disease.

Another example concerns the parameterization of the disease model:

There is not yet a consensus model of smallpox. We have designed a model that captures many features on which there is widespread agreement and allow us to vary poorly understood properties through reasonable ranges. (Eubank et al. 2004 , p. 183)

What "reasonable" means in this context, and how much it is related to available data, remains unclear. Finally, here is an example concerning the parameterization of the TV intervention:

Every simulated day, if contact tracing is in effect, a subset of the people on the list [of people showing symptoms] is chosen for contact tracing. […] In the experiment reported here, we use the fraction 0.8 and set the absolute threshold at either 10,000 or 1000. These are probably unrealistic numbers, but they allow us to estimate the best case results of a targeted vaccination strategy. (Eubank et al. 2004 , Supplement, 11)

In all of these examples, the very detail-demanding parameter and variable set poses the question of how to fill them with content. By default, one might assume that they are filled with empirical data. But it turns out that for these examples, empirical data is not available, or of too low a quality. So the modelers instead resorted to *ad hoc* assumptions, best guesses, or reasonable ranges.

I do not intend these observations as criticisms of the particular smallpox model, or of MSMs more generally. It seems perfectly reasonable to improvise on some parameters of one's model. But when discussing model choice, and in particular how to choose the resolution of detail of one's model, one should be mindful of how this choice affects the need to improvise. Imagine an extreme case, where a detail-poor model with only a few parameters that all can be determined from high-quality data can be developed into a detail-rich model, whose parameters can be filled only by *ad hoc* assumptions, best guesses, or use of reasonable ranges. Because these

improvisations carry a large chance of error, the detail-poor model is likely more accurate and therefore preferable for policy purposes than the detail-rich model. My MSM vs. ASM case is much less clear-cut than this extreme case, firstly because the parameters of the ASM are typically determined in a haphazard way, too, and secondly because the MSM does include a lot of certified data. However, there is a similar trade-off as in the extreme case, and that trade-off might in some cases lead to the conclusions that the ASM is a more powerful policy tool than the MSM.

4.3. Number of Parameters

Assume that the measurement of parameters was not a problem, so that 4.2 would not impose any constraints on the amount of detail incorporated in an MSM. In that case, another argument against such unchecked increase of detail arises from the comparative performance of such models in parameter estimation or calibration.

Disregarding technical detail, estimation and calibration both aim to determine values of unobservable model parameters by fitting the model to observable data. In the smallpox case, many parameters of the underlying TRANSIMS and EpiSims models are thus determined. To put it simply, the model takes census data, transport network data, land use data, etc. as inputs, and gives as output contact incidences, duration, and locations between individual agents. In accord with the generative program in simulation studies (Epstein 1999), model parameters are then adjusted so as to generate that model result that fits best with observational data. Once a close enough fit to such data has been achieved, the model is considered validated, and counterfactual policy interventions are introduced.

At first sight, MSMs appear to be better equipped to perform well in estimation or calibration exercises. If the target is of high complexity (which, in the case of vaccination policies, it undoubtedly is), then the more constraints one imposes on the model (in terms of the nature and number of its parameters), the less well such a model can fit the target. Conversely, the fewer constraints are imposed on a model, the better it can fit its target. Thus, it seems that MSMs can achieve a better fit to their targets than ASMs, and therefore appears as the more powerful policy tool.

The above intuition, although correct, misses an important trade-off that is well known in the model-selection literature. Although models with more free parameters have a larger potential to fit the target well, the larger number of free parameters often yields a lesser fit than the one achieved by a model with fewer parameters.

This trade-off becomes clearer by distinguishing two steps in the process of fitting a model to data. The first step consists in selecting a model—i.e., in specifying the number of parameters. Here, increasing the number of parameters indeed increases the model's *potential* to accurately represent the target.

The second step consists in calibrating or estimating the parameters based on a data *sample* drawn from the population. Increasing the number of parameters increases the model's fit to the sample—but this is not the ultimate goal. Rather, increasing the model's fit *to the target* is. Fitting the model "too closely" (i.e., by including too many parameters) to the sample will pick up on the inevitable random error in the sample, and thus leads to an increase in the divergence between model and target. This phenomenon is well known as "overfitting" in the statistics and machine-learning literature, and it applies to simulation modeling as well (Myung 2000).[2]

Selecting the right number of free parameters thus is the problem of "finding an appropriate compromise between these two opposing properties, *potential* and *propensity to underperform*" (Zucchini 2000, p. 45). As various studies have shown, if the sample size is large, adding more parameters above a certain threshold will not substantially increase fit to target; if sample size is medium or small, adding more parameters even decreases fit to target (Zucchini 2000; Gigerenzer and Brighton 2009).

This general finding also applies to the choice between MSM and ASM. In Sect. 2, I defined MSMs as containing many more parameters than ASMs. Consequently, MSMs are more subject to the danger of overfitting, and therefore more likely to fit the underlying target badly. Of course, whether in a particular case of comparing an MSM and an ASM the trade-off will favor one or the other is an open question (in particular, this is also the case for the two smallpox models, as a numerical study of their respective fit is beyond the scope of this paper). However, this general tendency makes it implausible to generally prefer MSMs over ASMs for policy purposes.

4.4. Number of Mechanisms

One of the important features of the simulation models discussed here is that they explicitly aim to represent *mechanisms*. In the smallpox case, both the MSM and the ASM were introduced as improvements over Kaplan's et al. (2002) macro model, because they explicitly modeled the population mixing mechanism instead of simply assuming homogeneous mixing. Nevertheless, the MSM and the ASM differ substantially in how they introduce such additional mechanisms. The smallpox ASM seeks to introduce a small number of simple mechanisms, while the MSM introduces a multitude of detail-rich mechanisms that are assumed to operate simultaneously.

In particular, the MSM distinguishes several activities at each location, each of which yields different contact rates; it also includes the effects of demographic factors (age in particular) on mixing; it distinguishes different forms of smallpox; and it tries to incorporate at least some rudimentary effects of infection on behavior. The ASM, in contrast, includes a lesser number of locations and does not distinguish activities or demographics; nor does it include infection effects on behavior.

Most observers seem to see the inclusion of additional mechanisms in comparison to the Kaplan et al. (2002) model as beneficial. It then also seems *prima facie* plausible to prefer the MSM to the ASM, as the former includes even more mechanisms and mechanistic detail than the latter.

Countering this intuition, I will use an argument made against the purported higher explanatory power of realistic simulation models. This argument has been put forward by Lenhard and Winsberg (2010), amongst others, with a specific focus on climate models. In short, they argue that with increasing complexity, models get more and more *opaque*; and this opacity prevents or at least reduces understanding the model components' contributions towards the model outcome.

More specifically, Lenhard and Winsberg argue that, with increasing complexity, the "fuzzy modularity" of a model increases. The more complex a model, the more subcomponents it has. Furthermore, when running a simulation on a complex model, these model components are run together and in parallel. But they do not all independently contribute to the model result. Rather, the components, in the course of a simulation, often exchange results of intermediary calculations among one another—so that the contribution of each component to the model result in turn is influenced by all those components that interacted with it.

The results of these modules are not first gathered independently and then only after that synthesized. […] The overall dynamics of one global climate model is the complex result of the interaction of the modules—not the interaction of the results of the modules [… D]ue to interactivity, modularity does not break down a complex system into separately manageable pieces. (Lenhard and Winsberg 2010, p. 258)

To put it differently, the effect of the multiple mechanisms is underdetermined more in an MSM than an ASM: first, due to the larger number of mechanisms included in an MSM, but also due to the increased interaction—the "fuzzy modularity"—of the mechanisms in the MSM. Clearly, there is more fuzzy modularity in a MSM like Eubank et al. (2004) than in an ASM like Burke et al. (2006). In the first place, this is a problem for the explanatory power of MSMs. Although MSMs might generate the *explanandum* quite closely, because of the higher degree underdetermination, it is more difficult in MSMs than in ASMs to infer from this fit which of the modeled mechanisms contributed to the generated result. If understanding consists in identifying the mechanisms that produced the *explanandum*, then a model's fuzzy modularity undermines improvements in our understanding.

This concern also applies to policy uses of MSMs. The model and simulation are supposed to help policymakers identify interventions that reliably produce desired outcomes in the relevant contexts. If we simulate such an intervention on a model that is severely underdetermined, then we don't know on which mechanisms (or interaction between mechanisms) the effect of the intervention was based. Something like this is the case in Eubank et al. (2004): their results might depend on some or all of the mechanisms in the model, or on their specific interaction, but it is impossible for the modelers to pry these influences apart. Such analyses are easier with ASMs, and for this reason they might be preferable for policy purposes.

4.5. Counterfactual Questions

By their nature, simulation studies for policy purposes involve counterfactual scenarios. In the models discussed here, this is the case at least in two instances: First, at the moment when the smallpox infection is introduced, and second, when the respective vaccination policies are implemented. For during neither of these modeling steps can the modeler point to actual data: there hasn't been a smallpox epidemic in an industrialized city in the twentieth century (that was not caught at the very early stage), nor have the different vaccination policies been tested in such environments. So even if the overfitting (4.3) and underdetermination (4.4) problems could be solved, modeling such counterfactual mechanisms cannot be validated by data fitting, because such data is not available.

Instead, parameters are set at some plausible values. Take, for example, the question of what fraction of identified people the TV intervention will likely be able to contact, and what the maximal capacity for such a program will be per day. Eubanks et al. cannot provide exact numbers, but instead suggest plausible values (see quote in Sect. 4.3). They then admit that "these are probably unrealistic numbers, but they allow us to estimate the best case results of a targeted vaccination strategy" (Eubank et al. 2004, Supplement, 11).

This poses the same problem for both MSMs and ASMs. However, in ASMs, this uncertainty about counterfactual mechanisms matches the uncertainty about the other components of the abstract model. Consequently, the policymaker is more likely to interpret the ASM model results as

possible outcomes that are affected by the uncertainty surrounding the mechanisms, variables, and parameters included in the model. The appearance of MSMs, in contrast, might propose a different interpretation: the uncertainty might appear to *dissipate* in the computational process, as multiple mechanisms and parameters that are interpreted as realistic interact with the uncertain components. Yet for reasons discussed in 4.4, it is typically not possible to discern by which components a model result was driven. Therefore, such a dissipation claim can typically not be sustained. Due to this opaque treatment of the uncertainty surrounding the counterfactual mechanisms, an ASM might often be preferable to an MSM for policy purposes.

4.6. Structural Uncertainty

From the discussion so far (as well as from common sense), it follows that uncertainty in model specification can never be fully eliminated, however little or much detail one might want to include in one's model. Some sources of uncertainty affect MSMs more than ASMs, as discussed in Sects. 4.2 and 4.3. But other inevitable uncertainties just stem from the general fallibility of human knowledge, and thus affect MSMs and ASMs equally. In this section, I will ignore the former differential problems and assume that MSMs and ASMs face the same degree of uncertainty. The question then is whether MSM and ASM offer different strategies for dealing with such inevitable uncertainty, and which of these strategies is better.

Consider the following example from Eubank et al. (2004). The contact data on which the simulation is based gives a detailed account of social interaction. The model lacks any account of how these social contacts may change under external shocks. The arrival of a threatening epidemic is, arguably, such a shock: it may well have important influence on how often people appear in public, go to work, or go to the hospital. The authors deal with this uncertainty as follows.

> One of the most important assumptions in any smallpox model is whether infectious people are mixing normally in the population. […] We undertook to model two (probably unrealistic) extreme cases: one in which no one who is infectious is mixing with the general population and another in which no one's behavior is affected at all by the disease. In addition, we modeled one more realistic case between these two extremes. (Eubank et al. 2004, Supplement, 11)

The model results strongly depend on the different assumptions. In particular, if people withdraw to the home, then all vaccination policies yield similar results, particularly if there is a delay in the vaccination procedure. However, if people do not withdraw, then LV is substantially less effective than either MV or TV (Eubank et al. 2004, p. 182, Figure 4).

Note that the MSM here only allows a qualitative distinction: depending on whether withdrawal occurs "early," "late," or "never," the simulation results in a different cumulative number of deaths. Such an analysis is similarly feasible with ASMs. The MSM authors do not assess the uncertainty included in these qualitative results beyond displaying them. While I agree that this seems the correct procedure in this case—as not enough evidence is available to provide a quantified assessment of the behavioral changes under shocks—the question then is why one would go through the additional efforts and costs of creating an MSM, if similar results could have been obtained with an ASM.

What MSMs often aspire to achieve instead is an overall quantification of the uncertainty involved. Although Eubanks et al. do not do this (correctly, I believe), they could have tried to specify a probability distribution over the different behavioral mechanisms and then represent the model

outcome as expected cumulative deaths. Such one-size-fits-all approaches in MSMs have been justly criticized for providing *false precision*:

f uncertainty is represented and reported in terms of precise probabilities, while the scientist conducting the analysis believes that uncertainty is actually 'deeper' than this—e.g. believes that available information only warrants assigning wide interval probabilities or considering an outcome to be plausible—then the uncertainty report will fail to meet the faithfulness requirement; it will have false precision. (Parker and Risbey 2015 , p. 4)

My argument here is that, in most applications of MSMs for policy purposes, non-quantifiable uncertainties arise. These should not be patched over by false precision, as described in the quote above. Alternatively, MSMs are used for providing different qualitative results, like the Eubanks et al. example above—which also could have been provide by an ASM. Defenders of MSMs here might reply that the advantage of such qualitative results from MSMs are more accurate than the comparative results from ASMs. However, my earlier arguments in Sects. 4.2–4.4 question whether this is necessarily the case. Consequently, the uncertainty quantification strategies facilitated by MSMs are not necessarily better than the strategy of ASMs.

4.7. Match with Decision Tools

MSMs, because they offer quantitative outcomes even when dealing with uncertainty, can easily be combined with standard quantified risk approaches to decision-making. For example, an MSM that gives possible outcomes of interventions (e.g., pairs of cost and cumulative deaths) at different probabilities can easily be combined with an expected value or an expected utility approach to decision-making. For this purpose, the outcome pairs are either monetized or their utility is determined, and this evaluation is weighted by the probability of this outcome. The policymaker chooses that intervention which yields the highest expected value or expected utility. ASMs, because they typically do not provide probabilities over uncertain outcomes, do not offer such a convenient procedure to the policymaker. For this reason, they might *at first* be considered inferior for policy purposes.

However, the combination of MSMs with quantified risk approaches is based on mere appearances, and lacks a justification. For reasons given in Sects. 4.4 and 4.6 , most models include inevitable uncertainty, which typically cannot be quantified. Furthermore, for reasons given in Sects. 4.2 and 4.3 , MSMs often include higher uncertainty than ASMs. Probabilistic quantifications of the uncertainty of MSMs, therefore, often represent claims that lack sufficient evidential support.

This leaves the question of whether MSMs could provide better support for qualitative decision approaches, which take uncertainties into consideration without quantifying them. These approaches include *structured qualitative decision-making*, which assumes that all relevant possible outcomes of an act can be identified, but that outcome uncertainty cannot be quantified; and *argumentative approaches* that identify only some of the relevant consequences (typically without being able to quantify their evaluation or their probability of occurring), while acknowledging that others are possible, too. Examples of the former include Maximin, Minimax regret and the O-P rule. Examples of the latter include pro-and-con tables and ordered checklists.

How would an MSM contribute in a better way to such approaches than an ASM? One answer is that we might be more confident in the possible outcomes that an MSM produces than in those of

the ASM, since the former includes more relevant detail both in terms of parameters and mechanisms. While I do not deny that this might be the case, such a conclusion is by no means necessary, as the MSM's outcomes are affected more by overfitting and underdetermination than the ASM's. Consequently, while both types of models typically support qualitative decision approaches, ASMs might occasionally better suited than MSMs, contrary to first appearances.

5. Conclusions

In this paper, I caution against an overly optimistic assessment of MSMs for policy purposes. In particular, I argued that MSMs might have more severe problems than ASMs in determining the accuracy of the model (4.2, 4.3); that MSMs might have more severe problems than ASMs in dealing with inevitable uncertainty (4.4, 4.6); and that MSMs might have more severe problems than ASMs with misinterpretation and misapplication due to their format (4.1, 4.5, 4.7). This of course does not exclude that some MSMs provide good justifications for policy decisions (and even better justifications than some ASMs); but it should caution against a general preference for MSM over ASMs for policy decision purposes.

References

Burke, Donald S., Joshua M. Epstein, Derek A. Cummings, Jon I. Parker, Kenneth C. Cline, Ramesh M. Singa, and Shubah Chakravarty. 2006. Individual-Based Computational Modeling Of Smallpox Epidemic Control Strategies. Academic Emergency Medicine 13 (11): 1142–1149.

Dawid, Herbert, and Giorgio Fagiolo. 2008. Editorial. Agent-based models for economic policy design: Introduction to the special issue. Journal of Economic Behaviour & Organization 67 (2): 351–354.

Epstein, Joshua M. 1999. Agent-based computational models and generative social science. Complexity 4 (5): 41–57.

Eubank, Stephen, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang. 2004. Modelling Disease Outbreaks In Realistic Urban Social Networks. *Nature* 429 (6988): 180–184. See supplement at http://www.nature.com/nature/journal/v429/n6988/extref/nature02541-s1.htm . Accessed 15 March 2016.

Farmer, J. Doyne, and Duncan Foley. 2009. The Economy Needs Agent-Based Modelling. Nature 460 (7256): 685–686.

Gigerenzer, Gerd, and Henry Brighton. 2009. Homo Heuristicus: Why Biased Minds Make Better Inferences. Topics in Cognitive Science 1 (1): 107–143.

Grüne-Yanoff, Till. 2011. Agent-Based Models as Policy Decision Tools: The Case of Smallpox Vaccination. Simulation and Gaming: An Interdisciplinary Journal 42 (2): 219–236.

Hartmann, Stephan. 1996. The World as a Process: Simulations in the Natural and Social Sciences. In Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View, ed. Rainer Hegselmann, Ulrich Mueller, and Klaus Troitzsch, 77–100. Dordrecht: Kluwer.

Hine, Dame Deirdre. 2010. The 2009 Influenza Pandemic: An independent review of the UK

response to the 2009 influenza pandemic. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/61252/the2009influenzapandemic-review.pdf . Accessed 02 May 2016.

Humphreys, Paul. 2004. Extending Ourselves: Computational Science, Empiricism, and Scientific Method. New York: Oxford University Press.

Kaplan, Edward H., David L. Craft, and Lawrence M. Wein. 2002. Emergency response to a smallpox attack: The case for mass vaccination. Proceedings of the National Academy of Sciences 99 (16): 10935–10940.

Lenhard, Johannes, and Eric Winsberg. 2010. Holism, entrenchment, and the future of climate model pluralism. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 41 (3): 253–262.

Myung, In Jae. 2000. The importance of complexity in model selection. Journal of Mathematical Psychology 44 (1): 190–204.

Parker, Wendy S., and James S. Risbey. 2015. False Precision, Surprise and Improved Uncertainty Assessment. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373 (2055): 20140453.

Schelling, Thomas. 1971. Dynamic models of segregation. Journal of Mathematical Sociology 1: 143–186.

Zucchini, Walter. 2000. An introduction to model selection. Journal of mathematical psychology 44 (1): 41–61.

[1] "We view the social networks created by TRANSIMS as a single instance of a stochastic process defined in an enormous space of possibilities" (Eubank et al. 2004 , Supplement, 3).

[2] This issue further compounds the problem of particular model targets when policy targets are more abstract, discussed in Sect. 4.1 . A close fit to the particular *model target*—even without the problem of overfitting—might not improve the model's usefulness for questions about the abstract *policy target*.