Synthese (2021) 199:3647–3673
https://doi.org/10.1007/s11229-020-02950-3

COMPUTATIONAL MODELLING IN PHILOSOPHY

The computational philosophy: simulation as a core philosophical method

Conor Mayo-Wilson[1] · **Kevin J. S. Zollman**[2]

Received: 13 March 2019 / Accepted: 3 November 2020 / Published online: 10 March 2021

Abstract
Modeling and computer simulations, we claim, should be considered core philosophical methods. More precisely, we will defend two theses. First, philosophers should use simulations for many of the same reasons we currently use thought experiments. In fact, simulations are superior to thought experiments in achieving some philosophical goals. Second, devising and coding computational models instill good philosophical habits of mind. Throughout the paper, we respond to the often implicit objection that computer modeling is "not philosophical."

Keywords Modeling · Simulations · Computation · Computational philosophy · Validation · Thought experiment · Philosophical methodology

Over the past several decades, computer simulations have made an inroad into philosophical work. Beginning with the pioneering work of Skyrms (1990, 1996, 2004, 2010) and Grim et al. (1998), many philosophers are incorporating computational models into their research. Computer simulations models are now making substantial

✉ Conor Mayo-Wilson
conormw@uw.edu

Kevin J. S. Zollman
kzollman@andrew.cmu.edu

[1] Department of Philosophy, University of Washington, Savery Hall, Room 361, Box 353350 , Seattle, WA 98195, USA

[2] Department of Philosophy , Carnegie Mellon University , Baker Hall 161 , Pittsburgh, PA 15213-3890, USA

appearances in social epistemology, ethics and political philosophy, philosophy of language, and philosophy of science.[1]

Although computer modeling is becoming more popular, it has not gained wide acceptance as a core philosophical method. Computer simulation is discussed in precisely one article in five recent handbooks dedicated to philosophical methodology;[2] The PhilPapers entry on "Philosophical Methods" mentions neither modeling nor simulation (Horvath 2019). Excepting the Munich Center for Mathematical Philosophy, we are aware of no graduate programs in philosophy that require a modeling or programming course. Some philosophers have recently taken aim at the value of the agent-based models most common in philosophy (Arnold 2014, 2015, 2019; Thicke 2019). Finally, as modelers, we can attest to hearing the following complaint time and again: "Interesting, but why is your research *philosophical*?"[3]

Modeling and computer simulations, we claim, should be considered core philosophical methods.[4] More precisely, we will defend two theses. First, philosophers should use simulations for many of the same reasons we currently use thought experiments.[5] In fact, simulations are superior to thought experiments in achieving some philosophical goals. Second, devising and coding computational models instill good philosophical habits of mind. Our second argument explains what *a modeler* learns from the act of modeling; the first explains what everyone can learn from computational models.

We were inspired to write this paper for two reasons. First, we think training philosophers in computational methods should be more common. Although we like logic, we think that logic should be one formal tool among many in philosophical reasoning. Modeling and programming are two important formal tools that fit naturally with paradigmatic philosophical methods.

Second, as modelers, we've encountered the same criticisms over and over again informally, in conferences, and in referee reports. Most frequently, we are simply told,

[1] The field is so large that we could not hope to capture all of it. But to get a sense for the diverse ways that simulations are employed here is a large sample. For examples in social epistemology see Betz (2013), Hartmann et al. (2009), Hegselmann and Krause (2002), Hegselmann and Krause (2006), Mayo-Wilson (2014), Zollman (2015). For ethics, social, and political philosophy see Alexander (2007), Bicchieri (2005), Bramson et al. (2017), Bruner and O'Connor (2016), O'Connor et al. (2019), Holman et al. (2018), Muldoon et al. (2012), Muldoon et al. (2014), Singer et al. (2017), Skyrms (1996), Skyrms (2004), Vanderschraaf and Skyrms (2003), Zollman (2005), Zollman (2008). For philosophy of language and logic see Franke and Correia (2017), Grim et al. (1998), Huttegger et al. (2010), O'Connor (2014a), O'Connor (2014b), Wagner (2009), Skyrms (2010). And for philosophy of science see Alexander (2013), Barrett (2007), Borg et al. (2018), Bruner (2013), Bruner and Holman (2017), Galeazzi and Franke (2017), Grim et al. (2013), Holman and Bruner (2015), Huttegger et al. (2015), Kummerfeld and Zollman (2016), Rosenstock et al. (2017), Smead (2010), Zollman (2007, 2010).

[2] Specifically, only (Weisberg 2016) discusses simulation. Simulation is not mentioned in Cappelen et al. (2016), D'Oro (2017), Daly (2015) and Haug (2013).

[3] One of us was asked this by a now colleague during a job interview.

[4] Although he does not specifically mention computer simulations, we think that the arguments in Williamson (2017, 2018) are consonant with ours regarding the value of modeling in philosophy. Our arguments differ in several respects from his, but they are not inconsistent.

[5] We do not wish to argue that simulations are a species of thought experiments, although we see the appeal of this way of describing things. For example, (Beisbart and Norton 2012) argue that simulations are more like arguments and not like traditional scientific experiments.

"Your model contains too many false assumptions to teach us anything of value." So in the last section of the paper—after we develop our argument for why simulations could be of use to philosophers—we collect and respond to the objections that we hear frequently. These objections are not entirely mistaken. Most are reasonable criticisms of *bad* simulations. So our goal is to use the objections to improve philosophical simulation. Throughout the paper, we respond to the often implicit criticism that computer modeling is "not philosophical."

Simulations can't help address every philosophical problem. No simulation will tell us whether abortion is moral. Moreover, simulations almost never answer philosophical questions by themselves. So simulations should not supplant other philosophical methods. Rather, simulations should be a tool in the philosopher's toolbox, to be used alongside thought experiments, careful analysis of arguments, symbolic logic, probability, empirical research, and many other methods. But for reasons we discuss below simulations are especially useful in several philosophical subfields, including social epistemology, social and political philosophy, and philosophy of science.

Section one contains our first argument. Philosophers have always used thought experiments, and we take it as given that thought experiments are an appropriate philosophical method. In Sect. 1.1, we describe six purposes of thought experiments. Our list is not exhaustive, and we make no attempt to address the rich philosophical literature on what thought experiments are (e.g., are they arguments?), how thought experiments are related to intuitions, and whether computer simulations and thought experiments are the same thing. By articulating the uses of thought experiments, however, we are able to argue for simulations by comparison. In Sect. 1.3, we argue that, for five of the six purposes that we identify, simulations are sometimes more effective than thought experiments.

Section two contains our second argument. While related, this argument is importantly distinct from the first. We describe several skills that philosophers prize: the ability to disambiguate claims, to recognize implicit assumptions in arguments, to assess logical validity, and more. We then explain how devising and programming computational models can foster those skills, *even if one has no intent of using the simulation results in construction of the final published argument.* Our claim is unusual in that it suggests that philosophers would benefit from using simulations privately as part of their argumentative development even if that doesn't ultimately show up in the finished product.

In the final section, we respond to some objections. These objections are not exhaustive, but they include the criticisms we hear most often from skeptics. We point out that with each objection comes an important lesson about how to simulations should be used in philosophical research.

1 Simulations and thought experiments

In this section, we defend the use of computational models in philosophical arguments. Our argument precedes by way of analogy to thought experiments. [6] As we are all

[6] [We are not the first to draw the connection between philosophical thought experiments and computer simulation Grim et al. (1998).]

familiar, philosophers often ask their readers to perform thought experiments and use the results of those for argumentative moves. We argue, in section Sect. 1.1, that philosophers often use thought experiments to achieve one of the following aims:

1. Elicit normative intuitions.
2. Justify counterfactual claims.
3. Explore logical relationships among philosophical theses.
4. Illustrate conceptual possibilities and impossibilities.
5. Distinguish explanatory reasons and identify those causes that explain a phenomenon.
6. Explore the dynamics of social and physical systems.

We don't claim this list is exhaustive, but rather that these represent several central ways that thought experiments are used. In Sect. 1.2, we reconstruct what we believe are the strongest arguments that thought experiments succeed in achieving these goals. Although we believe the conclusions of the arguments in Sect. 1.2, those conclusions are strictly not necessary for our argument.

We then argue in Sect. 1.3 that computer simulations can be—for the last five of these purposes—more effective than thought experiments. If computer simulations can achieve the same ends more effectively than traditional thought experiments, they then should be employed by the philosophical community.[7]

Only the most narrow interpretation of philosophy—that which equates philosophy with a specific method—could justifiably exclude computer simulations, and such an interpretation would rule out a wide swath of research that is typically called "philosophical." We address that last possibility in Sect. 1.4.

1.1 Thought experiments: six aims

The first use of thought experiments is perhaps most familiar: to evoke normative intuitions. Unhooking the violinist is morally justified. Pushing an innocent person onto train tracks is not. And so on.

We don't know what normative intuitions are, and we are agnostic about whether such intuitions are reliable. We mention this first use of thought experiments by way of contrast. Although cultivating intuitions might be the most salient use to some readers, thought experiments have been used many other ways, and arguably, the other uses are more common historically.

Philosophers use thought experiments to justify counterfactual claims, often when a real experiment is impossible, unethical, or impractical. In *Groundwork of Metaphysics of Morals*, for example, Kant asks us to imagine whether everyone could break promises when convenient. He concludes that, in such a world, no one would believe "promises" (Kant 2012) thus destroying the act of promising.

David Lewis (1969) defines conventional behavior in a thoroughgoing counterfactual way. To be a convention, a common behavior must have an alternative which could have been adopted. In the US we drive on the right side of the road but could have

[7] Importantly, we think that that *some* philosophers ought use simulations. This does not entail that every philosopher working on a problem should use simulations. We are firm believers in the division of cognitive labor and this applies as much in philosophy as it does in other domains.

driven on the left. Because of this, Lewis would call our practice of driving on the right a "convention." Other conventions require more imagination. Are the standards of logic conventional, as Carnap suggests? To answer that question, we must imagine how those standards might have been different.

For a last example of counterfactual reasoning, we turn to a core question in social epistemology: when are we justified in trusting others? Channeling Donald Davidson, Coady (1992) claims we're entitled to trust others by default because most human utterances must be true. Coady argues that, otherwise, utterances would not be understood as meaningful reports about the world.

> [I]magine a world in which an extensive survey yields no correlation between reports and facts . . . Imagine a community of Martians who . . . have a language which we can translate . . . with names for distinguishable things in their environment and suitable predicative equipment. We find, however, to our astonishment that whenever they construct sentences addressed to each other in the absence (from their vicinity) of the things designated by the names . . . they seem to say what we . . . observe to be false. But in such a situation there would be no reason to believe that they even had the practice of reporting.

We chose the above examples because we think readers will agree they contain squarely "philosophical" counterfactual claims. If the reader thinks the boundaries between philosophy and science are fuzzy (as we do), then examples can be multiplied almost indefinitely. Galileo asks us to imagine what would happen if a perfectly smooth ball were rolled on a frictionless "plane" that extended indefinitely around the Earth (Galilei 1967, pp. 147–148, 22). Such a "plane" would in fact be a spherical shell, and without friction, Galileo claims, the ball would orbit the Earth in perfectly circular motion. In general, thought experiments are used to justify counterfactual claims about not only people and societies, but also rotating buckets (e.g. in Newton), arrows (e.g., in Zeno and Lucretius), detached hands (e.g., Kant), and more.

The third use of thought experiments is related to the second: to explore logical relationships and to show that particular conclusions do *not* follow from common assumptions. Such thought experiments are sometimes called "destructive" Brown and Fehige 2017. Jarvis Thomson's (1971) violinist, for example, might show that the conclusion "It's unethical to kill a fetus" does not follow from the assumptions that "A fetus is a person" and "Fetuses are innocent of wrongdoing." Gettier (1963) cases are intended to show that "*S* knows that *p*" does not follow from the assumptions that *p* is true and that *S* justifiably believes *p*.

Fourth, thought experiments are used to distinguish explanatory reasons and to identify which "variables" explain a phenomenon. To dramatize the difference between "doing harm" and "allowing harm", for example, Foot (1967) compares two thought experiments. In the first, a judge frames a man to save five others, and in the second, a trolley driver flips a switch so that a runaway trolley kills one person not five. The judge is unethical; the driver is not. And the difference, says Foot, is explained by the fact the judge does harm, whereas the driver merely allows harm to be done.

As a last example, Danto (1983) imagines a gallery with completely identical red canvasses hung on the wall, each with a very different history: some accidental, others

intentional. Some are art, Danto argues, and others are not. He uses this to illustrate that nothing about the visual experience can explain what counts as art.

Fifth, thought experiments are used to illustrate possibilities and impossibilities. Hume, for instance, imagines someone who has never seen a particular shade of blue but is shown a color spectrum with the relevant missing shade. Hume admits that the subject might be able to imagine the missing shade. Hume's thought experiment is part of an admission that it's possible that not all simple ideas originate in simple impressions.

The final use of thought experiments that we'll discuss is often overlooked: to explore the dynamics of social and physical systems. Galileo routinely employs thought experiments concerning falling objects. To motivate the theories of special and general relativity respectively, Einstein imagines light clocks on trains and light beams passing through elevators. Importantly, these thought experiments ask us to imagine *motion, movement, or change*.

Squarely "philosophical" thought experiments often also involve imagining motion or change. For instance, to refine his counterfactual theory of causation, Lewis (1986) imagines two rocks are fired at a glass bottle but one strikes the bottle first. In general, debates about actual causation are full of thought experiments involving motion and collisions of physical objects.

Philosophical thought experiments often require us to imagine *social* dynamics, not just physical ones. Three examples discussed above—from Kant's *Groundwork*, Lewis' *Convention*, and Coady's *Testimony*—illustrate this point. For example, Kant asks us to imagine how people would react to changes in norms concerning promise-keeping. The dynamical nature of these thought experiments is sometimes hidden because we are asked to imagine a social system in *equilibrium*. For instance, Kant's thought experiment requires us to fast forward through the process of the dissolution of the norm of promise keeping and to imagine social interactions in a world in which the institution of promise-keeping has evaporated.

Perhaps the most widespread use of thought experiments about social dynamics is in the social contract tradition. Hobbes famously concludes that life without a sovereign would be "solitary, poor, nasty, brutish, and short" (Hobbes 1994, Chapter XIII).[8] Hume (1751, Section 3.1) asks us to imagine what justice would look like if one group in society were capable of completely dominating another. Nozick (1974) uses his famous "Wilt Chamberlain" thought experiment to argue that egalitarian societies will, through morally permissible wealth transfers, end up inegalitarian.

In short, many thought experiments are used to explore how societies would function in conditions that differ radically from our own and in conditions that may have never existed.

Again, the above list of uses of thought experiments is not exhaustive. There are also obvious relationships between the various uses of thought experiments; to illustrate a possibility (the fourth use), for example, is to illustrate a particular type of logical relationship among theses (the third use). Further, philosophical thought experiments often are used in multiple ways. But we think it's important to distinguish uses of

[8] Some scientists and philosophers have used computer simulations to explore philosophical claims made about the dynamics of anarchy (Martinez Coll 1986; Vanderschraaf 2019).

thought experiments to illustrate that talk of "intuitions" is sometimes too imprecise to distinguish good from poor uses of thought experiments. Kant's thought experiment, for example, might elicit the intuition that a world without promise-keeping would be bad or undesirable. But that *normative* intuition should be distinguished from Kant's *counterfactual* "intuition" that promise-keeping would fail to exist in societies in which promises were broken when convenient. The latter intuition, if it ought to be called "intuition" at all, is a claim about complex social systems, and it is amenable to empirical and mathematical investigation in ways the former normative intuition might not be.

1.2 What makes thought experiments successful?

Not all thought experiments are successful, but many of the examples from the previous section are often thought to be. Why?

Mach argues that thought experiments about *the mechanics of physical objects* are often reliable because they allow us to make use of implicit, non-propositional physical knowledge. He writes:

> Everything which we observe imprints itself uncomprehended and unanalyzed in our percepts and ideas, which then, in their turn, mimic the process of nature in their most general and most striking features. In these accumulated experiences we possess a treasure-store which is ever close at hand, and of which only the smallest portion is embodied in clear articulate thought. The circumstance that it is far easier to resort to these experiences than it is to nature herself, and that they are, notwithstanding this, free, in the sense indicated, from all subjectivity, invests them with high value. Mach (1883, p. 36). Quoted in Gendler (1998, p. 414)

Gendler (1998, 2004) expands upon Mach's reasoning, arguing that thought experiments are often useful because they allow us to reason with non-propositional representations, typically images. Gendler argues that mental manipulation of images employs psychological processes different than those used in deductive reasoning, and such processes are often essential for *producing a belief* in some proposition about the imagined objects or events. Although Gendler and Mach's arguments are controversial,[9] we grant their conclusions for the sake of our argument. Our question is, "Assuming Mach and Gendler's arguments are sound, which *types* of thought experiments are *reliable* for the purposes described in the previous section, and why?"

We think that Mach and Gendler's arguments most plausibly support the conclusion that visualization is useful for illustrating possibilities and logical relationships among various theses. Here, we expand on their arguments, drawing on work in philosophy of mathematics on diagrammatic reasoning (Giaquinto 2016; Shin et al. 2018).

In Euclidean geometry, a basic question is: which shapes can be drawn with only a straightedge and compass? At first, it might seem impossible to bisect an angle or construct a regular pentagon using these limited tools. But with the help of mental visualization and pen-and-paper, we can do a shocking amount.

[9] See Norton (2004) and Brown (2004) for alternative views.

Consider the construction of a square pictured in Fig. 1. In working through straightedge and compass constructions like this one, most people imagine the process: they engage in the thought experiment of construction. Instead of drawing a sequence of diagrams, we could have described the construction steps verbally. (*Begin with a line and two points A and B on that line. Draw a circle of arbitrary radius with center at B* …) But to *foster* and *justify* the belief that a square can be constructed from a line segment, this would have been less useful and, ultimately, would have required imagining the steps or actually implementing them with pen and paper to understand. Why?

Figure 1 is *easier to remember* than a sequence of verbal construction commands.[10] This makes it easier for a reasoner to revisit earlier parts of a long argument, which many philosophers since Descartes (at least) have recognized is required for a reasoner to have a justified belief in the conclusion.[11]

Further, Fig. 1 is *surveyable*. Checking whether a geometric diagram is a straightedge and compass construction is relatively easy. In a glance, one can see the construction utilizes only the relevant tools. With a bit more effort, one can be sure that the resulting diagram satisfies the definition of a square.[12]

Figure 1 is also mentally *manipulable*; we can re-imagine various parts of the diagram at will. In a glance, one can see that the distance between the first two points is arbitrary. So is the orientation of the first line, e.g. it could have been at a 45-degree angle relative to the page. With a little imagination, you can also see which parts are not arbitrary. For example, you can imagine what the resulting figure would be if the circles in Steps 5 and 6 had different radii.

Figure 1 is manipulable because it *omits* and *distorts*. It omits the precise distances and radii. It also distorts the lines, curves, and points, picturing them as thin but nonetheless two-dimensional objects.

Finally, diagrams allow us to reason geometrically even when we lack explicit propositional knowledge; this is a feature of thought experiments that Gendler and Mach emphasize. Almost everyone knows what a line and a circle is, even if they can't define it in set-theoretic language (i.e. that a circle is the set of all points in a plane equidistant from a given point).

Arguably, nearly everything we said about the Fig. 1 applies to Galileo's thought experiment about the ball on the frictionless plane. Our mental image of a ball on a plane can be recalled at will; it is surveyable because it involves two simple objects (a ball and plane), and it is manipulable: we imagine balls of different sizes, colors, and most importantly, material compositions behaving in exactly the same way. Finally, the thought experiment, as Mach argues, allows us to make use of our implicit knowledge of motion, which might be non-propositional.

[10] There is extensive psychological evidence linking visualization to memory, e.g., Cohen et al. (2009).

[11] Descartes (1984, p. 14, margin 370–371), "[V]ery many facts which are not self-evident are known with certainty, provided they are inferred from true and known principles through a continuous and uninterrupted movement of thought . . . deduction in a sense gets its certainty from memory."

[12] Proving that each side is of equal length is not complicated, but proving that steps 2-4 produce right angles is a bit more subtle (Euclid 1908, proof of Proposition 12, Book I).

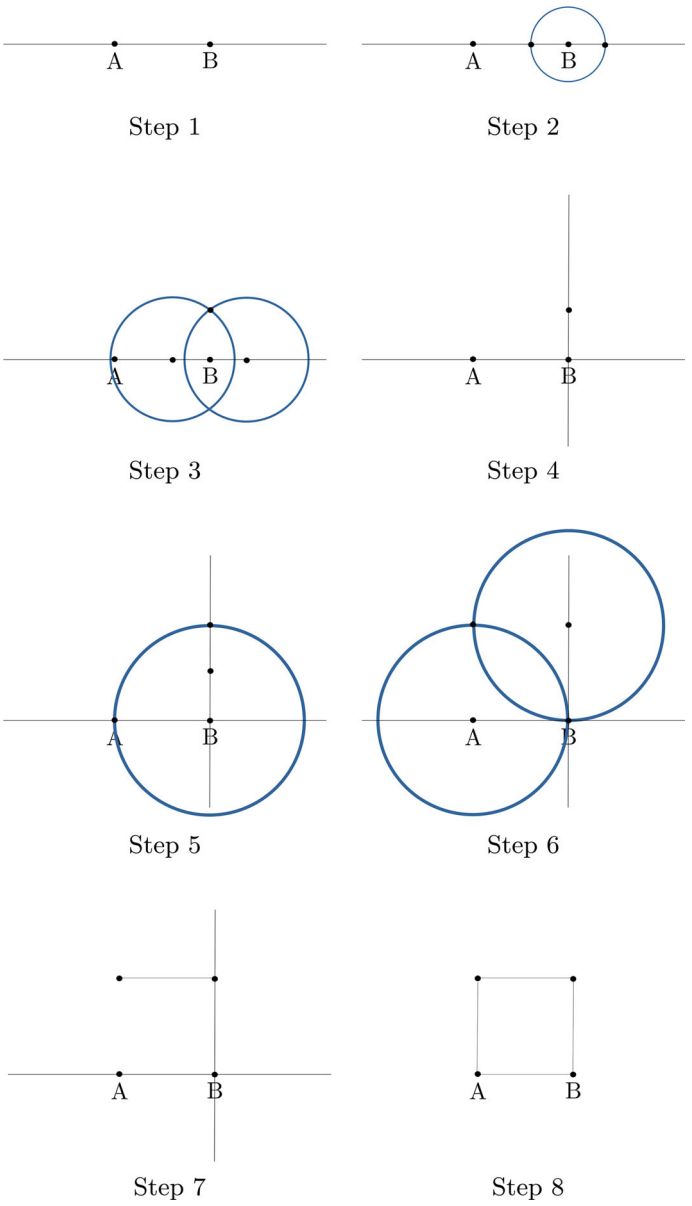

Fig. 1 Straightedge and compass construction of a square with arbitrary side length

Of course not all thought experiments involve visualizing physical systems. But the general point remains, if thought experiments are useful, it is likely because they engage parts of our cognition that are not propositional. By engaging these parts of cognition, an author hopes that a thought experiment will help in the construction, analysis, or recollection of philosophical arguments. So while our example here has

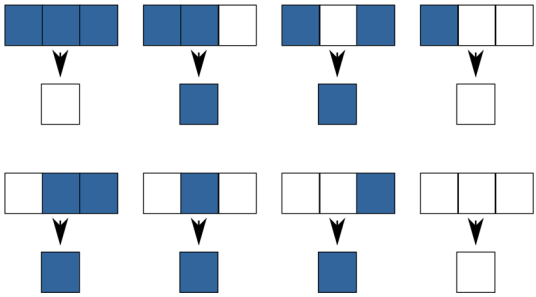

Fig. 2 A graphical illustration of Rule 110. The focal cell is pictured in the middle of a row of three. Shaded (blue) cells are "on" and unshaded (white) cells are "off." The single cell beneath each row indicates the next state of the focal cell. These eight transition rules fully define how every cell in an arbitrary array of cells evolves over time conditional on the state of its two neighbors. (Color figure online)

focused on the visual aspects of thought experiments, analogous virtues might be found for *some* of the other thought experiments described above.[13]

1.3 Simulations and thought experiments

We now argue that, when answering a philosophical question requires understanding the *dynamics of social systems*, simulations are better than corresponding philosophical thought experiments. Although we focus on social systems, many of our arguments apply equally well to physical systems involving multiple interacting bodies. To illustrate the usefulness of simulations in achieving the six goals enumerated in Sect. 1.1, we offer three examples.[14]

In 2013, the CDC asked polio researchers a very specific counterfactual question: what would have happened in the 2010 polio outbreak in Tajikistan if a larger age range of children had been vaccinated? Three distinct groups tackled the problem by building simulation models. Their simulations addressed the CDC's question with a high level of exactness: each simulation predicted how many additional people would have been saved by greater vaccination. Because human behavior influences how diseases spread, researchers *needed* models to explore social dynamics. The results of one such model is discussed in Wassilak et al. (2014), who found that the intervention would have had almost no positive effect on the outbreak, a somewhat counter-intuitive result.

A second famous (and infamous) example comes from an area known as cellular automata.[15] An old philosophical question is: what is the relationship between the complexity of a whole and the complexity of its parts? While this question can be made precise in many ways, the cellular automata Rule 110 provides a stunning illustration (see Fig. 2).

[13] It maybe that some philosophical thought experiments fail because they attempt to engage cognitive processes that we don't have or ones that are unreliable. This will—in effect—be our argument in favor of simulation for social systems that we present in the next section.

[14] We have chosen three examples from outside philosophy (or on the boundary between philosophy and other sciences) because we think the critical distance from philosophical problems will better illustrate our point to a philosophical audience. We could have chosen any of a number of wonderful philosophical simulations.

[15] Cellular automata lie on the boundary between computer simulation and mathematical modeling. This particular result was subject to a lawsuit regarding intellectual property, which has made it famous in two ways.

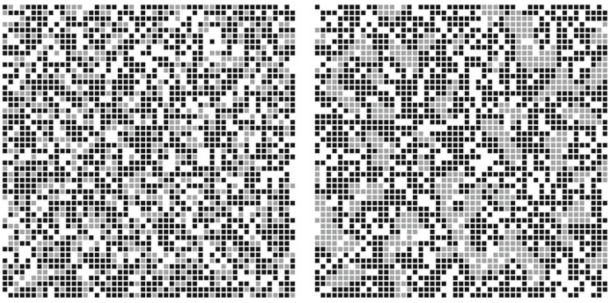

Fig. 3 An illustration of the starting and final state for a particular instance of the Schelling model. Here there are two types of agents, black and grey. Each agent will be unhappy if her type represents less than one-third of her neighbors and will move. These preferences lead to a highly segregated society where people are, on average, similar to three-quarters of their neighbors

Imagine a collection of cells, arranged in a line. Each cell has two states: on and off. There is a common clock, and with each tick, each cell updates its state based on the state of each of its neighbors (the cells on the immediate right and left). There are many rules that could govern the transition from one state to another, and Rule 110 is one such rule. There is nothing intuitively appealing about the rule, but it has one extremely important property: it is Turing complete (Cook 2004). That means that any computer program, no matter how complicated, could be implemented using a line of sufficiently many cells programmed to follow Rule 110. This shows it is possible to get almost arbitrary complexity out of something incredibly simple.

Rule 110 represents both purposes two and three of our enumeration: exploring the conceptual space by showing the connection between the simplicity of parts and the complexity of aggregate behavior.

Our final example is a famous model attributed to Schelling (1971).[16] The causes of segregation in modern cities are legion and well known. Institutional, explicit, and implicit discrimination make it impossible for people of certain races, nationalities, or religions to live in certain parts of a city. Once established, homogeneous neighborhoods often stay that way, and even once some of the more overt mechanisms are removed, segregation remains.

Schelling suggested another possible cause: perhaps a slight preference about the race of one's neighbors could also produce large-scale segregation. He imagined individuals arranged on a checker board, who would move if they were among a minority (e.g. less than 33%) in their neighborhood. Allowing time for relocation, this model produces large-scale segregation without any of the overt discrimination that features in the history of most cities (see Fig. 3). This shows that there might be important causes of segregation, that may be far more difficult to fight, than the institutional ones that feature so prominently.

Schelling's model achieves several of the goals outlined above. It identifies an important, possible cause of a critically important phenomena. It does so by exploring

[16] Schelling was anticipated in some ways by Sakoda (1971). For a complete discussion see Hegselmann (2017). Given that this model is universally known as Schelling's model, we will follow the convention of attributing it to Schelling.

a certain type of counterfactual—where there is no explicit discriminatory policy—regarding a complex social system.

Of course, we could go on. The social sciences are replete with examples of mathematical models and simulations achieving these various ends. But why do we think simulations are more reliable when social dynamics are concerned? Thought experiments, some argue, are successful in part because they require us to visualize a situation or event. Because some mental images are (i) easy to remember, (ii) surveyable, and (iii) manipulable (because they omit and distort), thought experiments might be effective tools for exploring the logical relationships among various philosophical and scientific principles. Further, mental images might also encode implicit, non-propositional knowledge that would be nearly impossible to use otherwise.

But, when thought experiments concern social systems, there is good reason to suspect our imagination is much less reliable. Social systems are complex in several ways that mechanical systems are not.[17]

Imagined social systems often contain more interacting agents than imagined physical ones. Kant asks us to imagine a society filled of people breaking promises, whereas Galileo asks us to imagine a single ball and plane. Further, the complexity and number of variables in imagined social systems is typically larger than that of mechanical systems; Galileo asks us to consider only the shape, weight, and speed of objects; Kant asks to consider the beliefs, desires, intentions, etc. of people.

Thought experiments about social *dynamics* are even more complicated. Equations governing the motion of mechanical objects are often geometrically representable (and so visualizable); the dynamical laws of social systems are typically not. Basic mechanical systems are more-or-less deterministic; most social systems are probabilistic.

Finally, our "implicit knowledge" of social systems is often not knowledge at all. Mach argues that our experience provides us with a wealth of mechanical knowledge, but his argument (if successful at all) relies on the fact that physical laws are constant across space and time. Our personal experience of the physical world, therefore, can be used to make inferences about the mechanics of objects at different times and places, and under conditions we have not encountered. But social norms vary widely around the world and across time; there's no reason to expect our local and recent experiences will help us understand the dynamics of societies with different norms, environments, and histories.

A surprising, clear illustration of this problem is provided by Wagner (2012). Although he does not describe it this way, Wagner's model provides a beautiful test bed for Coady's thought experiment. Imagine an alien species arrives. Since we have no language in common, we must learn to communicate with them about some matter of grave importance. To start let's suppose that we have a reoccurring interaction with the aliens where we are trying to develop a language with only two simple words, "true" or "false." We display some visual fact about the universe to the aliens and they respond with one of two prespecified words. We must come to learn which of those two words means "true" and which means "false."

[17] An anonymous referee points out that we say of "mechanical systems" is true of inclined planes but not of clouds and other complex physical systems. We agree, and the difference between clouds and planes provides further support for our argument: no contemporary scientist thinks it is reasonable to predict the behavior of clouds without computer simulations.

So far this describes a very simple version of the signaling game invented by Lewis (1969). If we suppose that the aliens want to communicate successfully with us and we want to communicate successfully with them, we will evolve to communicate effectively.[18] But Coady's situation is different: what if they don't want to communicate with us? What if our interests are completely opposed: they want us to believe false things and disbelieve true ones? Could it even be the case that we establish a system of communication with them where we could cogently say "everything they say is false?"

Take a moment and reflect on what you think about this situation, as a test for your intuitions. We already know what Coady's are. Have you decided? In the case with two predicates—"true" and "false"—your intuition was probably right. No meaningful language would exist between us and the aliens.

But what if we change the story in the most minor way? What if we introduce three potential predicates? Maybe now we want to discuss the location of something relative to another and we want to know is it much further away, much closer, or approximately the same distance. And suppose that, again, the aliens want to deceive us in a particular way. When the object is approximately the same distance, they want us to believe its further; when the object is further away, they want us to believe it's closer; and when the object is closer, they want us to believe it's approximately the same distance.[19]

Our intuition was that this small difference would make no difference, no meaningful language would exist. And, in one very strange sense, it's true. But the dynamics were nothing like what we imagined. Wagner (2012) shows that, under one model of learning, you have *chaos*.

"Chaos" is a term of art, and Wagner's paper argues that this case meets those conditions. For us, what matters is that the aliens will deceive us for a while, but then we will catch on, and then things will change again. But what's important about chaos is that those change points are completely unpredictable, even in theory. If you have the slightest error in your understanding of the current state of communication, you will be unable to predict whether or not the aliens or the humans will have the upper hand after some amount of time. Lest one think that Wagner's model is unusual, chaotic systems have been found in the study of other philosophically significant systems like the Prisoner's dilemma (Glance and Huberman 1993; Nowak and May 1992; Suzuki and Akiyama 2008).

Notice how shockingly simple Wagner's social system is. There are only two homogeneous groups: humans and aliens. They communicate using only three predicates. Yet the system is *in principle* unpredictable. Why expect that our intuitions are reliable for states of anarchy, like those imagined by Hobbes? Or for the dynamics of complex languages, like those imagined by Kant?

[18] There are many details that our short description leaves out. For one form of social learning, modeled by the replicator dynamics, this was first shown by simulation by Skyrms (1996) and then by more general mathematical proof by Huttegger (2007). For another model of learning called reinforcement learning, simulations are first discussed in Skyrms (2006) and mathematical proofs by Argiento et al. (2009). Of course there are other ways to model individual learning, some have been discussed others not (for a somewhat dated survey, see Huttegger and Zollman 2011).

[19] This is sometimes called cyclic dominance, and has a familiar structure from the children's game Rock-Paper-Scissors. If you played a lot of Rock-Paper-Scissors as a child, don't dismay, it turns out to be critical in evolutionary game theory (Szolnoki et al. 2014).

When complex social systems are at issue, simulations can be used to overcome these deficiencies of thought experiments. Simulations can be used to track the interactions of thousands of agents whose many features are governed by complex probabilistic laws. Purportedly "implicit knowledge" might likewise be encoded into a simulation, but unlike a thought experiment, one's "knowledge" is made *explicit* and *public*. It is, therefore, capable of being criticized, refined, and altered not only by the modeler but also by those who want to interpret the modeler's results and use them for their own purposes.

Computational models also inherit many of the virtues of thought experiments. To be of any use, computational models must omit features of the target system they represent; they often contain idealizations and distortions as well. And omissions, idealizations, and distortions in computational models have the same benefits they do when incorporated as parts of thought experiments. They allow one to isolate the important variables in explaining a social phenomenon; to explore whether the spread of a norm or the evolution of a particular behavior is *possible* under particular circumstances, etc.

Simulations are also "manipulable" like thought experiments. Just as details of thought experiments can be changed to test the robustness of a conclusion or the relationship between assumptions and conclusions, so can the code of model be updated and altered to check for robustness.

Finally, simulations of social systems are sometimes *visualizable*, even when a corresponding thought experiment would produce no concrete mental image. By rendering the agents and their properties in particular ways, simulations can make complex patterns—and dynamics in particular—available to the eye in a way that a thought experiment might not.

In short, for purposes of our argument, we grant that some thought experiments allow us to access *some types* of implicit, non-propositional knowledge. We also grant that such knowledge might not be incorporated into computational models. But those two admissions do not entail that thought experiments should be preferred, in all cases, to simulations. Why? We have argued that humans often do not have reliable implicit, non-propositional knowledge of social dynamics—or that such knowledge can be reliably distinguished from mere opinion and prejudice—and we have argued that simulations inherit many of the virtues of reliable thought experiments in precisely these circumstances.

1.4 When should a method become a core part of philosophy?

Even if simulations are better than thought experiments for achieving some philosophical ends, do the philosophical ends justify the computational means?

Neither ethical nor practical concerns speak against simulation. Typically the opposite is true. Like thought experiments, simulations can substitute for real experiments that are unethical, costly, or impractical. Given that philosophical thought experiments often involve remote conceptual possibilities and ethically grey scenarios, simulations

seem like an ideal way of exploring the social dynamics that philosophers would otherwise need to speculate about.[20]

But perhaps philosophers should leave model-building to scientists and engineers. Understanding climate models is obviously important for some ethicists and philosophers of science. But that doesn't mean that ethicists ought to learn how to construct climate models. A division of labor is necessary.

Unfortunately, unlike empirical research on climate change, certain philosophical questions are simply not being addressed by scientists. Further, a complete division of labor is typically impossible. Climate change ethicists need more than a passing familiarity with climate models. In general, to answer many philosophical questions, we philosophers might need to be able to manipulate existing models developed by scientists. Finally, if *all philosophers* lacked the ability to develop computational models, our community would be unable to interpret and evaluate scientific models that are relevant to our own work.

The last point clarifies why we've claimed only that the philosophical *community* should contain modelers. We do not claim that all philosophers should be modelers, even in cases in which models are indispensable. Societies need doctors, but no particular person must be a doctor. Similarly, philosophy needs computational modelers, but not all philosophers must be computational modelers. In fact, both empirical work and theory—including theoretical models developed by philosophers—suggest that philosophy benefits from a *diversity* of research approaches. One reason is that simulations and thought experiments are unreliable in different ways. Thus, philosophers, we think, should use both methods so as to discover and avoid errors associated with each method, in the same way that two scientific methods might be used to estimate a quantity, even if one is believed to be more accurate in the case at hand.

Finally, critics might grant simulation is a fine research method but not a *philosophical* one. "Simulation is just not philosophy," one might say. Such a critic either confuses a descriptive for a normative claim or begs the question entirely. We grant that, historically, computer simulation *has* been rare in philosophy (as it has been in every field!). Our thesis concerns methods that philosophers *ought* to use more often. And to baldly assert that "Simulation *ought not* be considered philosophy" is just to beg the question.

2 Simulations and philosophical habits of mind

We now argue that modeling and programming foster philosophical habits of mind. This argument is distinct from that in Sect. 1 which focused on how a philosophical argument might benefit from including simulation models as part of the argument. In this section, we argue that *modelers* benefit from developing and programming computational models, *even if their models are never read by others*. Just as many philosophers sketch their arguments in logical or pseudo-logical notation to check

[20] This is not to say there are no ethical questions about conducting thought experiments or simulations. Like any form of research there are important ethical questions about fraud, transparency, inductive risk, etc.

for validity, developing a simulation can force one to uncover hidden assumptions or ambiguities that would go unnoticed without such an exercise.

Many of the skills that modelers develop, we believe, correspond to the five uses of models and thought experiments we focused on in the previous section. That is, if our arguments in the previous sections were successful, then modelers should (just as readers who assess philosophical arguments containing models) become more successful at justifying counterfactual claims, exploring logical relationships among philosophical theses, developing concrete descriptions of "possibility spaces", distinguishing explanatory reasons, and exploring the dynamics of social and physical systems. Thus, in this section, we focus on two additional philosophical skills that, we think, are especially advanced by devising and programming models, namely, the skills of (i) identifying implicit assumptions in arguments and (ii) disambiguating claims and distinguishing concepts. We then argue that the benefits of modeling typically outweigh the harms and that no other method is known to be as effective in acquiring some philosophical skills.

We take inspiration from one of Josh Epstein's arguments to the social science community:

> The first question that arises frequently – sometimes innocently and sometimes not – is simply, "Why model?" …my favorite retort is, "You are a modeler." Anyone who ventures a projection, or imagines how a social dynamic – an epidemic, war, or migration – would unfold is running some model.
>
> But typically, it is an implicit model in which the assumptions are hidden, their internal consistency is untested, their logical consequences are unknown, and their relation to data is unknown. But, when you close your eyes and imagine an epidemic spreading, or any other social dynamic, you are running some model or other. It is just an implicit model that you haven't written down (Epstein 2008).

Ultimately, our argument rests on an empirical assumption, namely, that constructing computational models helps one acquire certain philosophical skills. We admit that our evidence for the premise is derived from personal experience and untested (but plausible) causal hypotheses. As modelers, we have ample first-hand experience of cases in which developing a model has clarified our own thinking and suggested fruitful paths for research. As teachers, we have seen students' philosophical thinking improve by developing computational models.[21]

[21] One can, and should, develop tests for our empirical premise. But as an argumentative matter, it is worth noting that philosophers accept the efficacy of philosophical training on the basis of equally weak evidence. Relatively little empirical research compares the effect of traditional philosophical training with that of training in other disciplines (e.g., in economics or mathematics). So it's also an empirical question whether teaching other core philosophical methods—like thought experiments—is the best way of teaching one to do philosophy.

2.1 Simulations promote real thinking

We now explain why, when investigating social dynamics, developing computational models helps the modeler practice (i) identifying implicit assumptions in arguments and (ii) disambiguating claims and distinguishing concepts.[22]

Let's start with the ability to identify implicit assumptions. Again, consider Kant's claim that we would stop taking promises seriously if everyone broke promises when convenient. If a philosopher were to develop a simulation model, they must ask and answer many more questions. To examine Kant's claim, a modeler must *represent* (a) actions like making, breaking, and keeping promises, (b) properties of agents, such as their beliefs (e.g., about how likely various people are to break promises) and their interests (so we can know what it means for breaking a promise to be "convenient" or in the agent's self interest), and (c) relationships among agents (e.g., with whom do agents most frequently communicate? Are some agents more likely to need to make promises and others more likely to need to decide whether to accept or deny promises?). Even at this early stage, the modeler is forced to ask questions that Kant simply never asks: should we represent beliefs by binary variables (Tom is reliable vs. not), qualitatively scaled items (e.g., Tom is very reliable, somewhat reliable, somewhat unreliable, etc.), or numerical variables (e.g., Tom's reports are true $x\%$ of the time)? What does it mean for an act to be in the agent's best interest? For example, is "best interest" captured by some expected utility model, a maximin principle, or something else entirely?

As these representational choices are made, a modeler is also forced to make *dynamical* assumptions: how do agents' beliefs, interests, and behaviors change over time? Here, again, the modeler must ask questions that Kant does not. What information do agents *learn* when a promise is broken? For instance, if Tom breaks a promise to Sally, does Sally learn so? Do others learn of that broken promise? If so, who? How much do people *remember*? How is an agent's behavior a function of her beliefs and desires? In short, a modeler is forced to answer dozens of questions that casual consideration of Kant's thought experiment would not require one to answer.

One might stop at this point and say: "these questions are irrelevant, Kant's claim doesn't depend on these." But how can we know without developing a model? Surprising results depend on very subtle assumptions about how people learn.[23]

Anyone who develops an explicit mathematical model (computational or not) of Kant's thought experiment should answer questions like the ones we've posed. But we suspect that trying to *program* a model makes the questions especially forceful. The reason is fairly simple: you can't hide assumptions from a computer. A computer doesn't know what your variables are supposed to represent. It won't draw semantic inferences from the names you use for variables. For instance, a computer won't assume a variable called "belief" represents something propositional. Further, a computer doesn't know what types of standard assumptions are made about belief updating or rational choice.

[22] We are not the first to notice the virtues of modeling a practice for the philosopher. Williamson (2017, 2018) also suggests that devising and manipulating models helps to instill important philosophical habits.

[23] These issues show up in philosophy where different assumptions about how individuals learn can have interesting implications on evolution of communication in Lewis signaling games (Barrett and Zollman 2009; Huttegger et al. 2010; Skyrms 2010).

Contemplating how to create a *computational* model of Kant's thought experiment, in short, forces one to identify the implicit assumptions about the nature of belief, rationality, social interaction, and so on, that Kant makes to reach the conclusion that promise-keeping would cease to exist in a world in which breaking promises "when convenient" was universal. In identifying those implicit assumptions, a modeler is then forced to draw distinctions (e.g., between different representations of belief) that a reader who engages with Kant's original text would not make.[24]

Of course, some choices made by the modeler will be arbitrary. In fact, every modeling assumption she might consider could be unrealistic. What is important is that, in constructing a model, she recognizes that various psychological, sociological, etc. assumptions are necessary *for any argument whatsoever*, even those that *purport* not to rely on modeling assumptions. To paraphrase Epstein, Kant was running some model or other. It's just not one that he wrote down.

2.2 Tradeoffs: skills versus bad habits

In devising and programming computational models, philosophers develop, practice, and hone their philosophical abilities. But does it follow that, ceteris paribus, philosophers should devise and program computational models?

As we discussed in Sect. 1.4, devising and programming computational models is not normally unethical, costly, or impractical. And although it's plausible that there are other methods that might help one hone one's philosophical abilities more effectively, we don't know of one more effective for philosophical questions involving social dynamics.

Nonetheless, one might worry that the intellectual benefits of computational modeling are outweighed by the bad habits it encourages. Some modelers, we hypothesize, might adopt non-robust assumptions for the sake of running a simulation of some type. Other modelers might adopt implausible assumptions simply because a particular programming language (e.g., NetLogo) makes those assumptions easy to implement.[25] By itself, adopting false or implausible assumptions is not a sin, for reasons we will discuss below. But becoming confident in the conclusions drawn from models constructed in these ways—or a collection of models sharing similar implausible assumptions—is a problem.

We grant that unreflective modeling can encourage some bad habits. But we remind the reader: *consider the alternatives*. The same is true for any method. Unreflective commitment to to a naive philosophical method can engender logically correct, but useless, philosophical argument—as anyone who has taught an undergraduate course can attest. The solution to both problems is not to abandon the method, but rather to

[24] During this process of identifying and disambiguating assumptions, a modeler is naturally led, just as a reader of such a model might be, to question whether the assumptions necessary for Kant's argument are reasonable and whether Kant's conclusion would follow from slightly different assumptions about how agents are represented and how their properties change. This later process is called "robustness" testing and in a frequent task completed by modelers in scientific domains (Odenbaugh and Alexandrova 2011; Weisberg 2006).

[25] This point has been made explicitly by the philosopher and modeler, Jason McKenzie Alexander in talks and personal conversation.

improve it and make the practitioner understand the limitations of their own research strategies.

Avoiding intellectual vice, whether those encouraged by modeling or more traditional philosophical methods, requires training and diligence. Just as we think philosophers can learn to avoid the above sins of demands for rigor, we think modelers can learn to express greater modesty when the best models are unfit for the desired purposes.

3 Objections

Before addressing objections, we note that simulation is now indispensable in science, from physics, to climate science and geology, to biology and social sciences. Without simulations there could be no modern science.[26] For this reason, some objections to modeling in philosophy would be equally applicable to scientific uses of computer simulation modeling. We're not suggesting that philosophers should just "trust the scientists." There are bad simulations in the sciences as well. But we urge the reader to consider the following: if an objection attacks the epistemic benefit of all computer simulations, the philosopher must be prepared to dispense with an enormous amount of successful scientific practice.

3.1 Your model is false

One might object, "your model is false" or "Your assumption that X is false."[27] Yes, we know. The important question is, "Do the false assumptions undermine the model's intended *purpose*?"

Consider again the uses of thought experiments and simulations that we discussed above. Some simulations, like thought experiments, are intended to show that a particular event, phenomenon, or dynamics is *possible*. For example, signaling games are often used to show how organisms might develop complex patterns of communication using only extremely simple learning strategies. The goal of such models is not to show that, for example, vervet monkeys *did* evolve to produce alarm calls in a particular way, but rather, that scientists might not need to postulate complex mental states in order to explain the development of a primitive language. Similar remarks apply to

[26] For philosophers of science, simulations raise many interesting epistemological questions. How can simulations tell us something about the real world? What distinguishes good simulations from bad ones? Are simulations always replaceable by some sort of analytic model? See (Downes 2021; Winsberg 2010; Weisberg 2013) for many of the issues.

[27] One of us once discussed a mathematical model from economics with a political philosopher, who remarked that the model couldn't possibly be relevant to their philosophical work. In philosophy, the political philosopher said, we are concerned with valid arguments with true premises. Since the model made false assumptions, it could not be part of an argument with true premises. This view of modeling is a natural consequence of the logical picture of modeling that accompanied the semantic view of theories in philosophy of science. Along with most contemporary philosophers of science, we believe this way of viewing modeling is the result of an equivocation of the way "model" is used in logic and in the various sciences (Downes 1992; Weisberg 2013). This objection can also be mistakenly raised in scientific contexts as well (Waldherr and Wijermans 2013).

models used to show how self-interested organisms might develop cooperative norms, even if such organisms frequently find themselves in competitive situations. Models intended to illustrate a conceptual possibility (e.g., to provide a "how possible" story) need not contain exclusively true assumptions.

Consider a second use of models: to identify important variables or distinctions. Mayo-Wilson (2014), for example, argues that philosophers should consider honest miscommunication and network structure when investigating when testimony is trustworthy. Mayo-Wilson's model is not intended to support a particular policy, but rather, to show that idealizations made in philosophical thought experiments are not harmless: variables that some social epistemologists ignore are often crucial for identifying when to trust others. When models are used in this way, again, it's not essential that all the model's assumptions are true.

The same is true of models that are intended to explore logical relationships among various theses, and in particular, to show that certain widely-held conclusions do not follow from common assumptions. For example, science-policy makers often assume that, as long as scientists are honest and truth-seeking, it is always beneficial to encourage scientists to share their findings and seek out others' work before continuing with their own research. Zollman (2007) shows that might not be the case. Even if one is skeptical about the robustness of Zollman's results (cf. Rosenstock et al. 2017) or thinks Zollman's idealizations are suspect, the model is still of value: it forces a policy-maker to ask the question, "Is there any reason the sharing of information might backfire in the case at hand?" Models intended to show that a conclusion does *not* follow from assumptions need not contain only true assumptions.

Even when a model is used to draw a conclusion *about the world*, however, it is often not fruitful to show that a particular modeling assumption is false. Planets aren't perfect spheres; planes aren't frictionless, and collisions among gas molecules aren't perfectly elastic. But physicists regularly make those assumptions and succeed anyway. We won't try to answer the question of which idealizations are useful (and when). We want only to emphasize the following. Some false assumptions are useful, and others aren't. The use of false assumptions is not *by itself* an objection to modeling practice as a whole.

The gap between a model and the real world is always inductive (Sugden 2000). Good models are like the real world in some respects and unlike it in others. When a model resembles the world in some respects X, it will, as a matter of contingent fact, turn out to resemble the world in other ways Y. But the correlation between X and Y is never discoverable *a priori*. So one ought not ask "Are the model's assumptions true?" but rather, "Does the model resemble the world in the relevant respects for the question at hand (Weisberg 2013; Waldherr and Wijermans 2013)?"

3.2 Your model isn't validated

Our critic might grant that false models are often useful. They might object, however, that most models developed by philosophers are not *validated*, i.e., the predictions of the models have not been tested against the real world (Martini and Pinto 2016; Thicke 2019). That is, the critic grants that many false models yield reasonable predictions.

For instance, models that describe the planets as point masses are fairly accurate for describing planetary motions over thousands of years. The critic just denies that philosophers' models are useful in this way.

Again, we emphasize that simulations, like thought experiments, have many purposes, and validation just isn't necessary for many of those purposes. For example, to illustrate that certain events or situations are *possible*, it's often not necessary to validate one's model.

Of course, it is important to validate models (when possible) for particular purposes. Validation is extremely important in epidemiology, for example. This is in part why the CDC asked the simulation teams to use the historical case of an actual polio outbreak. This allowed the modelers to choose parameters that fit the actual outbreak and only alter the one variable.

But even when validation is important, it can't always be achieved. The outbreak of COVID-19 occurred during this revision of this paper, and early models of the progression of that disease were often wrong. But we could not wait for careful validation of the predictions of those models before using them for policy intervention. Doing so would have left millions dead. As of the writing of this paper, now months into the pandemic, we may still not yet be in a position to have carefully validated models of the disease.

In such cases, scientists and philosophers use the word "validation" to mean something like, "testing the *assumptions* of the model against the world." To validate an early model of COVID-19, for example, one might ask "is the modeled disease transmitted at the same rate that we think COVID-19 is? Is the model of the way people move and interact realistic?" Further validation might involve checking the model's parameter settings against the world or against another validated model. As we are all aware, even models that have be carefully validated in this more indirect way can be wildly off. But, sometimes that is the best that one can do.

Unfortunately, there are times where validation, even in this weaker sense, is complicated or impossible. One should be honest about these limitations, but it is not always a reason to abandon modeling entirely. It's not that we should trust, uncritically, an unvalidated model. But rather that a weak inductive argument, properly understood, is better than no argument at all. Most importantly, the cases that make model validation impossible will make it likewise impossible to assess the (typically implicit) assumptions of arguments that do not use an explicit model.

Finally, when comparing computational models to other forms of argumentation, it's important to distinguish apparent from actual validation. Leading scientific journals like *Nature* and *JAMA* routinely publish short science-policy proposals. Such proposals often contain quantitative empirical data and basic statistical models (e.g., regressions). Scientists then use this data to defend counterfactual claims about how science would be if we adopted a novel policy.

In contrast, agent-based models of science (in both philosophy and science) are rarely motivated by quantitative empirical data. Instead, such models are often justified by "plausibility" arguments and stylized historical case studies. So on first glance, the former statistical models seem better validated than the latter agent-based ones.

But care is needed. All science policy proposals rely on causal hypotheses about how scientific institutions, corporations, and individual scientists respond to incentive

schemes. Philosophers' agent-based models make those causal hypotheses explicit. Two-page editorials in *JAMA* rarely do. And the statistical models published in science policy papers almost never justify the required causal conclusions. Instead, we conjecture, the implicit causal assumptions are accepted unconsciously on the basis of qualitative plausibility arguments and observations of current scientific practice. With regard to science policy proposals, *all existing arguments would benefit from validation*, but we don't see any reason to suppose the causal hypotheses implicit in scientists' reasoning will be easier to validate than those in philosophers' models.

3.3 Your model might have a coding error

Computer programs can contain errors. So occasionally, a computational model represents a system different from the one the modeler intended.[28] One might object that this possibility represents a reason to exclude modeling from philosophical discussion.

As before, we ask the reader to consider the alternatives. In order for coding errors to be a reason to abandon simulation methods, it would need to be the case that coding errors were *more* common than other forms of conceptual errors like equivocation, fallacious reasoning, and the like. And we see no reason to suppose that.

But even if coding errors were extremely common, we have already illustrated an important benefit of computer simulation. Modelers should make their code available to others. When they do so, errors are uncovered and fixed. So even if coding errors were more common than other forms of error—something we do not believe—simulations would not be epistemically inferior because those errors that sneak through would be easier to detect and remedy. And even if code is not published, scholars often attempt to replicate models—sometimes finding critical errors when they do (Will and Hegselmann 2008).

Like the other objections, there is an important grain of truth behind this one. Computer scientists have developed powerful methods for detecting coding errors. Philosophers should be trained to use these methods. With such training, we can reduce a source of error without abandoning a fruitful method.

3.4 Epistemic opacity

In principle, one can check mathematical proofs step-by-step. Similarly, a rigorous philosophical argument contains all the steps necessary to reach a conclusion.[29] In principle the reader has all the relevant facts to reconstruct the justification for the claims offered in a paper using one of these methods. For example, we provided you with all you need to see how to construct a square using a straightedge and compass in Fig. 1.

Humphreys (2008) argues that, in contrast, simulations are "epistemically opaque" because even experts may not be able to see how the results were generated. Of course,

[28] This possibility lies behind the dialog among Alexander et al. (2015), Thoma (2015), Weisberg and Muldoon (2009).

[29] Note that *in practice* published mathematical proofs often contain large gaps that specialists must fill in and published philosophical arguments are notorious for their enthymematic structure.

a simulator could provide the code, but that may not be enough to fully understand how or why the simulation produced particular results.

Humphreys argues that opaqueness is a unique feature of computer simulations. We disagree: experiments are often analogously opaque. A scientist can provide you with her "raw" data, but those data do not provide you with any information about how or why the experiment produced a given result. Further, raw data are often the output of some detectors, the observations of a lab assistant, etc., and one who lacks knowledge of the data gathering procedures will likewise lack knowledge of why the experiment produced a given result.

Similarly, although the opacity of simulations stands in contrast to the visual thought experiments we discussed in Sect. 1.1, not all thought experiments are epistemically transparent. We often cannot explain why some actions are ethical or why some objects count as pieces of art. A virtue of thought experiments is that they allows us to use non-propositional knowledge that may be opaque to us. In this respect, some thought experiments are more opaque than computer simulations.

Humphreys does not think epistemic opacity is a reason to exclude simulations from scientific practice. However, if one wanted to use this concept in order to argue against philosophical simulations (but not a scientific ones) then one would need to argue that philosophers should be more concerned about epistemic opacity than scientists. Perhaps such a argument is possible, but we are hard pressed to devise one that isn't question begging. Why should philosophy be more epistemically transparent than the sciences? And if it should be, why are thought experiments acceptably opaque when simulations are not?

Here too, there is grain of truth behind the objection: computer simulations can be more or less transparent. A modeler can describe her model perfectly, summarize her simulation results in excruciating detail, and yet, fail to explain why the model produced the given results.

That is bad practice, but it's not an inherent limitation of simulation models: some modelers make the relationship between model and results crystal clear. Philosophers of science should characterize what makes those modelers successful.

Social and professional norms can help as well. If journals require modelers to make their code available, reviewers can ask for additional explanation or analysis to help make the results of models less opaque. Improving philosophical training, therefore, can help to make simulation practice better by creating a pool of reviewers and editors who know the right questions to ask.

4 Conclusion: the computational philosophy

Leibniz once said, "Calculemos!" (Let us calculate!) We say, "Let us simulate."

Our computational philosophy resembles, in some important ways, the mechanical philosophy of Locke, Galileo, and Leibniz among others. Just as the mechanical philosophers were skeptical of *a priori* speculation about the physical world, we are skeptical of *a priori* speculation about the social world. So just as the mechanical philosophers urged that the methods of philosophy be extended to include controlled

experiments (of sometimes artificially simple physical systems), we urge philosophers to embrace simulations of social dynamics.

But unlike the mechanical philosophers who deemed certain methods and types of explanations unintelligible, we're pluralists about philosophical methodology. So we end with a thought experiment that, we think, cannot be replaced by a social simulation. In upcoming centuries, human brains might be augmented by digital computers that allow us to remember and compute in ways that we currently cannot. Imagine future philosophers can, without any effort, mentally run social simulations by accessing computers that have been implanted in their brains; sometimes those philosophers run simulations unconsciously. Would the computationally-augmented "thought" experiments of future researchers count as philosophy? We think so, and we see no reason to think that current simulations run "outside" the brain are any less philosophical.

References

Alexander, J. M. (2007). *The structural evolution of morality*. Cambridge: Cambridge University Press.

Alexander, J. M. (2013). Preferential attachment and the search for successful theories. *Philosophy of Science*, *80*(5), 769–782.

Alexander, J. M., Himmelreich, J., & Thompson, C. (2015). Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science*, *82*, 424–453.

Argiento, R., Pemantle, R., Skyrms, B., & Volkov, S. (2009). Learning to signal: Analysis of a micro-level reinforcement model. *Stochastic Processes and their Applications*, *119*(2), 373–390.

Arnold, E. (2014). What's wrong with social simulations? *The Monist*, *97*(3), 359–377.

Arnold, E. (2015). How models fail. In C. Misselhorn (Ed.), *Collective agency and cooperation in natural and artificial systems, number 122 in philosophical studies* (pp. 261–279). New York: Springer.

Arnold, E. (2019). Validation of computer simulations from a Kuhnian perspective. In C. Beisbart & N. Saam (Eds.), *Computer simulation validation* (pp. 203–224). New York: Springer.

Barrett, J. A. (2007). Dynamic partitioning and the conventionality of kinds. *Philosophy of Science*, *74*(October), 527–546.

Barrett, J. A., & Zollman, K. J. (2009). The role of forgetting in the evolution and learning of language. *Journal of Experimental and Theoretical Artificial Intelligence*, *21*(4), 293–309.

Beisbart, C., & Norton, J. D. (2012). Why Monte Carlo simulations are inferences and not experiments. *International Studies in the Philosophy of Science*, *26*(4), 403–422.

Betz, G. (2013). *Debate dynamics: How controversy improves our beliefs*. New York: Springer.

Bicchieri, C. (2005). *Grammar of society*. Cambridge: Cambridge University Press.

Borg, A. M., Frey, D., Šešelja, D., & Straßer, C. (2018). Epistemic effects of scientific interaction: Approaching the question with an argumentative agent-based model. *Historical Social Research/Historische Sozialforschung*, *43*(1), 285–307.

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., et al. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science*, *84*(1), 115–159.

Brown, J. R. (2004). Peeking into Plato's heaven. *Philosophy of Science*, *71*(5), 1126–1138.

Brown, J. R., & Fehige, Y. (2017). Thought experiments. In Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy*.

Bruner, J. P. (2013). Policing epistemic communities. *Episteme*, *10*(04), 403–416.

Bruner, J. P., & Holman, B. (2017). Experimentation by industrial selection. *Philosophy of Science*, *84*(December), 1008–1019.

Bruner, J. P., & O'Connor, C. (2016). Power, bargaining, and collaboration. In T. Boyer, C. Mayo-Wilson, & M. Weisberg (Eds.), *Scientific collaboration and collective knowledge*. Oxford: Oxford University Press.

Cappelen, H., Gendler, T. S., & Hawthorne, J. (Eds.). (2016). *The oxford handbook of philosophical methodology* (1st ed.). New York, NY: Oxford University Press.

Coady, C. A. J. (1992). *Testimony: A philosophical study*. Oxford: Oxford University Press.

Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences*, *106*(14), 6008–6010.
Cook, M. (2004). Universality in elementary cellular automata. *Complex Systems*, *15*, 1–40.
Daly, C. (Ed.). (2015). *The Palgrave handbook of philosophical methods*. New York: Palgrave-Macmillian.
Danto, A. (1983). *The transfiguration of the commonplace*. Cambridge: Harvard University Press.
Descartes, R. (1984). *The philosophical writings of descartes* (Vol. 1). Cambridge: Cambridge University Press.
D'Oro, G. (Ed.). (2017). *The Cambridge companion to philosophical methodology*. Cambridge, UK: Cambridge University Press.
Downes, S. M. (1992). The Importance of Models in theorizing: A deflationary semantic view. In *PSA: Proceedings of the Biennial meeting of the philosophy of science association* (pp. 142–153).
Downes, S. M. (2021). *Models and modeling in the sciences: A philosophical introduction*. New York: Routledge.
Epstein, J. M. (2008). *Why model?* Technical report 4, Santa Fe Institute.
Euclid (1908). *The thirteen books of Euclid's elements*. Cambridg: Cambridge University Press.
Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 1–7.
Franke, M., & Correia, J. P. (2017). Vagueness and imprecise imitation in signalling games. *The British Journal for the Philosophy of Science*, *69*(4), 1037–1067.
Galeazzi, P., & Franke, M. (2017). Smart representations: Rationality and evolution in a richer environment. *Philosophy of Science*, *84*(3), 544–573.
Galilei, G. (1967). *Dialogue concerning the two chief world systems, Ptolemaic and copernican*. California: University of California Press.
Gendler, T. S. (1998). Galileo and the indispensability of scientific thought experiment. *The British Journal for the Philosophy of Science*, *49*(3), 397–424.
Gendler, T. S. (2004). Thought experiments rethought—and reperceived. *Philosophy of Science*, *71*(5), 1152–1163.
Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.
Giaquinto, M. (2016). The epistemology of visual thinking in mathematics. In Zalta, E. N. (Ed.), *The stanford encyclopedia of philosophy* (Winter 201 ed.). Metaphysics Research Lab, Stanford University.
Glance, N. S., & Huberman, B. A. (1993). The outbreak of cooperation. *The Journal of Mathematical Sociology*, *17*(4), 281–302.
Grim, P., St Paul, H., Mar, G., & Denis, P. S. (1998). *The philosophical computer*. Cambridge: MIT Press.
Grim, P., Singer, D. J., Fisher, S., Bramson, A., Berger, W. J., Reade, C., et al. (2013). Scientific networks on data landscapes: Question difficulty, epistemic success, and convergence. *Episteme*, *10*(4), 441–464.
Hartmann, S., Martini, C., & Sprenger, J. (2009). Consensual decision-making among epistemic peers. *Episteme*, *6*(2), 110–129.
Haug, M. (Ed.). (2013). *Philosophical methodology: The armchair or the laboratory?*. London, New York: Routledge.
Hegselmann, R. (2017). Thomas C. Schelling and James M. Sakoda: The intellectual, technical, and social history of a model. *Journal of Artificial Societies and Social Simulation 20*(3).
Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence. *Journal of Artificial Societies and Social Simulation*, *5*(3).
Hegselmann, R., & Krause, U. (2006). Truth and cognitive division of labor: First steps toward a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation*, *9*(3), 1–29.
Hobbes, T. (1994). *Leviathan*. Indianapolis: Hackett Publishing Company.
Holman, B., Berger, W. J., Singer, D. J., Grim, P., & Bramson, A. (2018). Diversity and democracy: Agent-based modeling in political philosophy. *Historical Social Research*, *43*(1), 259–284.
Holman, B., & Bruner, J. P. (2015). The problem of intransigently biased agents. *Philosophy of Science*, *82*(5), 956–968.
Horvath, J. (2019). Philosophical methods. https://philpapers.org/browse/philosophical-methods. Accessed on 18 Nov 2020.
Hume, D. (1751). *An enquiry concerning the principles of morals*. London: A. Millar.
Humphreys, P. (2008). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615–626.
Huttegger, S. M. (2007). Evolution and the explanation of meaning. *Philosophy of Science*, *74*(January), 1–27.

Huttegger, S. M., Bruner, J. P., & Zollman, K. J. (2015). The handicap principle is an artifact. *Philosophy of Science*, *82*(December), 997–1009.
Huttegger, S. M., Skyrms, B., Smead, R., & Zollman, K. J. (2010). Evolutionary dynamics of Lewis signaling games: Signaling systems vs. partial pooling. *Synthese*, *172*(1), 177–191.
Huttegger, S. M., & Zollman, K. J. (2011). Signaling games: The dynamics of evolution and learning. In A. Benz, C. Ebert, G. Jäger, & R. van Rooij (Eds.), *Language, games, and evolution*. Berlin: Springer.
Jarvis Thomson, J. (1971). A defense of abortion. *Philosophy and Public Affairs*, *1*(1), 47–66.
Kant, I. (2012). *Groundwork on the metaphysics of morals*. Cambridge: Cambridge University Press.
Kummerfeld, E., & Zollman, K. J. (2016). Conservatism and the scientific state of nature. *British Journal for the Philosophy of Science*, *67*(4), 1057–1076.
Lewis, D. (1969). *Convention: A philosophical study*. Cambridge: Harvard University Press.
Lewis, D. (1986). *Philosophical papers: Volume II*. Oxford: Oxford University Press.
Mach, E. (1883). *the science of mechanics, La Salle, IL* (6th Ed.). Open Court Original.
Martinez Coll, J. C. (1986). A bioeconomic model of Hobbes "state of nature". *Social Science Information*, *25*(2), 493–505.
Martini, C., & Pinto, M. F. (2016). Modeling the social organization of science. *European Journal for Philosophy of Science*, *7*(2), 221–238.
Mayo-Wilson, C. (2014). The reliability of testimonial norms in scientific communities. *Synthese*, *191*(1), 55–78.
Muldoon, R., Lisciandra, C., Bicchieri, C., Hartmann, S., & Sprenger, J. (2014). On the emergence of descriptive norms. *Politics, Philosophy and Economics*, *13*(1), 3–22.
Muldoon, R., Smith, T., & Weisberg, M. (2012). Segregation that no one seeks. *Philosophy of Science*, *79*(1), 38–62.
Norton, J. D. (2004). On thought experiments: Is there more to the argument? *Philosophy of Science*, *71*(5), 1139–1151.
Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *259*, 826–829.
Nozick, R. (1974). *Anarchy, state, and utopia*. New York: Basic Books.
O'Connor, C. (2014a). Ambiguity is Kinda good sometimes. *Philosophy of Science*, *82*(1), 110–121.
O'Connor, C. (2014b). The evolution of vagueness. *Erkenntnis*, *79*(S4), 707–727.
O'Connor, C., Bright, L. K., & Bruner, J. P. (2019). The emergence of intersectional disadvantage. *Social Epistemology*, *33*(1), 23–41.
Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: Robustness analyses in economics and biology. *Biology & Philosophy*, *26*(5), 757–771.
Rosenstock, S., Bruner, J. P., & O'Connor, C. (2017). In epistemic networks, is less really more? *Philosophy of Science*, *84*(2), 234–252.
Sakoda, J. M. (1971). The checkerboard model of social interaction. *The Journal of Mathematical Sociology*, *1*(1), 119–132.
Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, *1*(2), 143–186.
Shin, S.-J., O. Lemon, and J. Mumma (2018). Diagrams. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 201 ed.). Metaphysics Research Lab, Stanford University.
Singer, D. J., Bramson, A., Grim, P. Holman, B., Jung, J., Kovaka, K.,Ranginani, A., & Berger, W.J., (2019). Rational social and political polarization. *Philosophical Studies, 176*(9), 2243–2267.
Skyrms, B. (1990). *The dynamics of rational deliberation*. Cambridge: Harvard University Press.
Skyrms, B. (1996). *Evolution of the social contract*. Cambridge: Cambridge University Press.
Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. New York: Cambridge University Press.
Skyrms, B. (2006). *Signals*. In *Presidential address, Philosophy of Science Associate Meeting*, Vancouver, B.C.
Skyrms, B. (2010). *Signals: Evolution, learning and information*. New York: Oxford University Press.
Smead, R. (2010). Indirect reciprocity and the evolution of "moral signals". *Biology & philosophy*, *25*(1), 33–51.
Sugden, R. (2000). Credible worlds: The status of theoretical models in economics. *Journal of Economic Methodology*, *7*(1), 1–31.
Suzuki, S., & Akiyama, E. (2008). Chaos, oscillation and the evolution of indirect reciprocity in n-person games. *Journal of Theoretical Biology*, *252*(4), 686–693.

Szolnoki, A., Perc, M., Mobilia, M., Jiang, L.-L., Szczesny, B., & Rucklidge, A. M. (2014). Cyclic dominance in evolutionary games: A review. *Journal of The Royal Society Interface*, *11*(100), 20140735–20140735.

Thicke, M. (2019). Evaluating formal models of science. *Journal for General Philosophy of Science*.

Thoma, J. (2015). The epistemic division of labor revisited. *Philosophy of Science*, *82*(3), 454–472.

Vanderschraaf, P. (2019). *Strategic justice: Convention and problems of balancing divergent interests*. Oxford: Oxford University Press.

Vanderschraaf, P., & Skyrms, B. (2003). Learning to take turns. *Erkenntnis*, *59*, 311–348.

Wagner, E. O. (2009). Communication and structured correlation. *Erkenntnis*, *71*(3), 377–393.

Wagner, E. O. (2012). Deterministic chaos and the evolution of meaning. *The British Journal for the Philosophy of Science, 63*(3), 547–575.

Waldherr, A., & Wijermans, N. (2013). Communicating social simulation models to sceptical minds. *Journal of Artificial Societies and Social Simulation*, *16*(2013), 13.

Wassilak, S. G., Pallansch, M. A., Duintjer Tebbens, R. J., Cochi, S. L., Kalkowska, D. A., & Thompson, K. M. (2014). The potential impact of expanding target age groups for polio immunization campaigns. *BMC Infectious Diseases*, *14*(1), 2.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, *73*, 730–742.

Weisberg, M. (2013). *Simulation and similarity*. Oxford: Oxford University Press.

Weisberg, M. (2016). Modeling. In H. Cappelen, T. S. Gendler, & J. Hawthorne (Eds.), *The oxford handbook of philosophical methodology* (1st ed., pp. 262–286). New York, NY: Oxford University Press.

Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, *76*, 225–252.

Will, O. and R. Hegselmann (2008). A replication that failed: On the computational model in 'Michael W. Macy and Yoshimichi Sato: Trust, cooperation and market formation in the U.S. and Japan. Proceedings of the National Academy of Sciences, May 2002'. *Journal of Artificial Societies and Social Simulation 11*(3), 1–24.

Williamson, T. (2017). *Model-building in philosophy* (pp. 159–172). New York: Wiley-Blackwell.

Williamson, T. (2018). Model-building as a philosophical method. *Phenomenology and Mind*, *15*, 16–22.

Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago: University of Chicago Press.

Zollman, K. J. (2005). Talking to neighbors: The evolution of regional meaning. *Philosophy of Science*, *72*, 69–85.

Zollman, K. J. (2007). The communication structure of epistemic communities. *Philosophy of Science*, *74*(5), 574–587.

Zollman, K. J. (2008). Explaining fairness in complex environments. *Politics, Philosophy, and Economics*, *7*(1), 81–97.

Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis*, *72*(1), 17–35.

Zollman, K. J. (2015). Modeling the social consequences of testimonial norms. *Philosophical Studies*, *172*(9), 2371–2383.