Getting Serious about Similarity

Michael Weisberg*†

Although most philosophical accounts about model/world relations focus on structural mappings such as isomorphism, similarity has long been discussed as an alternative account. Despite its attractions, proponents of the similarity view have not provided detailed accounts of what it means that a model is similar to a real-world target system. This article gives the outlines of such an account, drawing on the work of Amos Tversky.

Philadelphia is a racially diverse city. According to the 2010 census, the population was about 44% African American, 39% Caucasian, 12.5% Latino, and 5.4% Asian. Nevertheless, when one looks more closely at census tracts, one sees a very typical urban demographic pattern: racial clustering by neighborhood. There are probably many factors driving this demographic pattern, but an interesting model of one possible factor was developed by Thomas Schelling in the 1970s (1978). He wanted to know if it was possible for a city to segregate even if its members did not have strong preferences to live in segregated neighborhoods.

To investigate this question, Schelling constructed an agent-based model consisting of two racial groups and a grid representing a city. In the original version, the model was physically instantiated on a chessboard, with dimes and nickels representing two types of individuals, *A* and *B*, and the squares on the chessboard representing spatial location. Apart from the individuals and their initially random spatial layout, the model also contained a utility function and a movement rule. The utility function said that each individual prefers that at least 30% of its neighbors be of the same type. So the *A*s want at least 30% of their neighbors to be *A*s and likewise for the *B*s. Schelling's

*To contact the author, please write to: Department of Philosophy, 433 Cohen Hall, University of Pennsylvania, Philadelphia, PA 19104; e-mail: weisberg@phil.upenn.edu.

†Many thanks to Matt Bateman, Brett Calcott, Zoltan Domotor, Alkistis Elliott-Graves, Emily Parke, Isabelle Peschard, Daniel Singer, Martin Thomson-Jones, Bas van Fraassen, Scott Weinstein, and Deena Skolnick Weisberg for comments and suggestions. This research was supported, in part, by National Science Foundation grant SES-0957189.

Philosophy of Science, 79 (December 2012) pp. 785–794. 0031-8248/2012/7905-0009$10.00

Figure 1. Schelling's model of segregation on a 51 × 51 grid with 2,000 agents. Each agent prefers 30% of its Moore neighbors to be the same shape and color. The initial distribution of agents was random, and the model equilibrated after 14 time steps.

neighborhoods were defined as standard Moore neighborhoods, a set of nine adjacent grid elements. An agent standing on some grid element *e* can have anywhere from zero to eight neighbors in the adjoining elements.

The movement rule can be described as follows: agents sequentially choose to either remain in place or move to a new location. When it is an agent's turn to make a decision, it determines whether its utility function is satisfied. If it is satisfied, the agent remains where it is. If it is not satisfied, then the agent then moves to the nearest empty location. This sequence of decisions continues until all of the agents satisfy their utility function.

The dynamics of a modern computer-based implementation of Schelling's model are shown in figure 1. This clearly shows that starting from a random distribution of agents, the equilibrium state of the model is segregation. Thus, Schelling's major result is that small preferences for similarity can lead to massive segregation. This result is quite robust across many changes to the model, including different utility functions, different rules for updating, differing neighborhood sizes, and different spatial configurations. In fact, it is extremely hard to avoid segregation when agents have some preference for like neighbors (Muldoon, Smith, and Weisberg 2012).

Schelling offered his model as a *how-possibly explanation*, one potential mechanism by which neighborhoods could segregate. But what if the mechanism described by this model is actually part of the explanation of how Philadelphia came to have the neighborhood structure that it has? If this were the case, in what relation to Philadelphia would this highly idealized model of the segregation dynamic stand? The answer, I believe, is that the model would be similar to Philadelphia. Highly idealized models are not truthful representations of their targets, nor are they isomorphic to their targets; they are similar to their targets. In this article, I will give a sketch of how a similarity-based account of the model/world relation can be developed.

1. Similarity. Although a number of philosophers have advocated similarity as the ideal candidate for the model/world relation (e.g., Cartwright 1983;

Giere 1988), similarity has a checkered history in philosophy, and appeals to it have seemed dubious to many philosophers. For example, W. V. O. Quine argued that any general notion of similarity is deeply problematic because we cannot explain it in terms of more empirically or logically basic notions. On this basis, he concluded that the concept of similarity is "logically repugnant" (1969, 59) and that mature sciences dispense with similarity relations altogether.

While Quine's argument primarily rested on simple chemical examples, he also invoked more foundational arguments from Nelson Goodman. One of Goodman's arguments against similarity is that appeals to it merely label something unknown, rather than giving a characterization of the relationship in question. A proper analysis ought to give a reductive definition of similarity, but, Goodman argues, no such definition exists (1972).

Another of Goodman's arguments is that similarity is too promiscuous of a relation. For any three objects, he argued, there will always be some respect in which two of the objects resemble each other more than the third. If we have a green square, a red square, and a red circle, there is no obvious pair whose members are more similar to each other than the third.

Goodman uses this second problem to show that there can be no context-free similarity metric, either in the trivial case or in a scientifically realistic case. Such arguments led philosophers like Giere and Cartwright to restrict their accounts of model/world similarity. Giere tells us that a model must resemble its target in certain "respects and degrees" (1988, 93), presumably given to us by background theory and a theorist's interests. Cartwright tells us that the relevant similarity between models and their targets is "behavioral similarity," which I interpret to mean similarity of causal structure (1983).

I think that Giere and Cartwright are on the right track here. Similarity does seem to be the relation that holds between models and the world because it comes in degrees, can be used to compare idealized models to targets, can relate qualitative features of models to targets, and so forth. However, they give us few details about what similarity supervenes on, how it depends on context, how similarity judgments are to be evaluated, and so forth. This article takes the first steps toward an account of this relation, beginning with the work of Amos Tversky.

2. Tversky's Contrast Account. In the 1970s, Tversky developed a set-theoretic account of similarity with which he tried to capture the everyday judgments of similarity and dissimilarity made by his experimental subjects. At the time, the most sophisticated theory of similarity judgments had been developed by Frank Attneave (1950) and Roger Shepard (1980, 1987), drawing on some of Quine's ideas. In Shepard and Attneave's *geometrical* account of similarity, objects are assigned to a location in multidimensional space on the basis of values assigned to their features. Similarity, then, is just

the distance between points representing objects in this space. For example, colors might be represented as coordinates in a three-dimensional space, corresponding to their lightness, hue, and saturation. Two colors could then be compared to each other by measuring the distance between them. The closer two objects are in this feature space, the more similar they are to one another.

Tversky thought that this was not a fully general account of similarity for a number of reasons. For one thing, he believed that not all properties relevant to similarity judgments can be mapped onto a dimension of a property space; some features are qualitative. He also believed that not all similarity judgments were symmetric. His subjects judged that North Korea was more similar to China than China was to North Korea. So Tversky wanted an account that was more flexible and general then the geometric account but that could also generate the results of the geometric account when they applied.

To a first approximation, Tversky's *contrast* account of similarity says that the similarity of objects a and b depends on the features they share and the features that they do not share. His account is developed in the following way: we begin with some set of features Δ called the feature set. These can be quantitative or qualitative predicates and might include elements such as "is red," "is to the left of X," "will land on heads with probability 0.5," and just about anything else. For two objects a and b, we will define A as the set of features in Δ possessed by a and B be the set of features in Δ possessed by b. Further, we choose some weighting function $f(\cdot)$, which is defined over $\mathcal{P}\Delta$. The similarity of a to b is then given by the following equation:

$$S(a, b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A). \quad (1)$$

For some set of features Δ; weighting function $f(\cdot)$; and term weights θ, α, and β, this equation will give us a similarity score that can be used in comparative judgments of similarity. It says that the similarity of a to b is a function of the features they share, penalized by the features that they do not share. Tversky thought that the term weights and weighting functions were context sensitive and that the rules governing their filling in would be discovered by empirical psychology (Tversky 1977; Tversky and Gati 1978).

I think that Tversky's contrast account makes an excellent starting point for developing a more formal similarity-based account of the model/world relation. The simplest possible version of such an account would be to swap out Tversky's generic objects a and b with models (m) and targets (t). If we did this, we could write down an equation of the following form:

$$S(m, t) = \theta f(M \cap T) - \alpha f(M - T) - \beta f(T - M), \quad (2)$$

where M and T are defined in the manner of A and B in equation (1).

To a first approximation, I think that this is the correct account of the model/world relationship. A model is similar to its target, or to a mathematical representation of its target, when it shares certain highly valued features and does not have many highly valued features missing and when the target does not have many significant features that the model lacks. Relevant features are identified in a natural or formal language, and their importance is weighted relative to the goals of the scientific community. In order to transform this basic idea into an account of the model/world relation, we need to consider in more detail where f, Δ, and the weighting coefficients come from. And we also need to attend to the form of the equation itself.

3. Attributes and Mechanisms. In scientific inquiry, it is typical to distinguish the properties and patterns of a system from the underlying mechanism that generates these properties. When a model is constructed, we can also distinguish among static properties, static patterns, dynamic patterns, and the processes that generate these patterns. I propose that we make a major division between properties and patterns, on the one hand, and the underlying generating processes, on the other. Call the first category *attributes* and the second category *mechanisms*. A more abstract way to think about the difference between attributes and mechanisms is that attributes are states and state transitions, and mechanisms are transition rules.

As an example of the distinction between attributes and mechanisms, consider equilibrium states of Schelling's segregation model. When the model comes to equilibrium, it contains racially segregated clusters, and it approaches this state with a pattern of "contagion," where small clusters lead to bigger clusters. What drives these patterns are the agents' utility functions and rules for movement. Attributes such as degrees of clustering are states of the model, and mechanisms such as agents' movement rules are the transition rules of the model. Insofar as Schelling's model explains segregation in actual cities, then there has to be some relation between the model's attributes and the city's attributes. And there has to be some relation between the model's transition rules and the actual mechanisms that drive segregation in the city.

With this division in mind, we can take the initial account of model/target similarity described in equation (2) and divide its terms for model and target features into two categories: attributes and mechanisms. I will designate these with the subscripts a and m. The expression for similarity becomes

$$\begin{aligned} S(m, t) = {} & \theta f(M_a \cap T_a) + \rho f(M_m \cap T_m) \\ & - \alpha f(M_a - T_a) - \beta f(M_m - T_m) \\ & - \gamma f(T_a - M_a) - \delta f(T_m - M_m). \end{aligned} \tag{3}$$

We now have expressions for the intersection of attributes ($M_a \cap T_a$) and mechanisms ($M_m \cap T_m$) as well as the difference between the model and target's attributes and mechanisms. Additionally, each of the six terms can now be weighted independently, which is an important part of the explanation of how theorists treat different kinds of models used for different purposes.

For the moment, assume that we adopt the simplest possible feature weighting function $f(\cdot)$, where each element of Δ is weighted equally. This is equivalent to saying that each term in the equation has the numerical value of its cardinality. We can also use the simplest possible weights for the individual terms. If we set $\theta = \rho = \alpha = \beta = \gamma = \delta$, we can just drop the weights from our expression. In this case, the equation becomes

$$\begin{aligned} S(m, t) = {} & |M_a \cap T_a| + |M_m \cap T_m| \\ & - |M_a - T_a| - |M_m - T_m| \\ & - |T_a - M_a| - |T_m - M_m|. \end{aligned} \tag{4}$$

Equation (4) makes the basic structure of our modified Tversky equation clear. A model is more similar to its target when it shares more attributes and mechanisms, and it is systematically penalized when the model contains extraneous detail and when it fails to capture or incorrectly captures features of the target. To standardize the scale between 0 and 1, with $s = 1$ equivalent to maximal similarity, we can write equation (4) as a ratio. Rather than taking similarity to be a measure of the features shared minus the features not shared, we can take it to be the ratio of features shared to those not shared. Further, if we normalize the equation, we can ensure that similarity values are bounded between 0 and 1, corresponding to maximally dissimilar and identical, relative to Δ.

Rewriting equation (4) in ratio form yields

$$S(m, t) = \frac{|M_a \cap T_a| + |M_m \cap T_m|}{|M_a \cap T_a| + |M_m \cap T_m| + |M_a - T_a| + |M_m - T_m| + |T_a - M_a| + |T_m - M_m|}. \tag{5}$$

Before reintroducing the coefficients and weighting function, let us consider how the different terms would get filled for the Schelling model and Philadelphia. The model was primarily aimed at generating shared attributes ($M_a \cap T_a$). It reproduced patterns of racial segregation, and choosing the correct utility function could create whatever racial exposure values were observed. However, it is almost certainly not the case that real populations have fully shared, simple utility functions ($M_m - T_m$). Similarly, even the most gridlike cities such as Philadelphia are not completely regular grids like in the Schelling model ($M_a - T_a$).

4. Feature Sets. Tversky was extremely liberal about the elements of Δ, and my further breaking up the set's contents into attributes and mechanisms does not impose much in the way of further constraint. Some of this liberalness is deliberate because of the myriad comparisons theorists are required to make. For example, the elements of Δ can be qualitative, interpreted mathematical features such as "oscillation," "oscillation with amplitude *A*," "the population is getting bigger and smaller," and so on. They can be strictly mathematical terms such as "is a Lyapunov function." Or they can be physically interpreted terms such as "equilibrium" or "average abundance."

Which of these kinds of terms should go in Δ? There is no context-free answer to this question. A combination of context and prior practice will determine how both model and system are broken up into parts and properties. This is conceptually before the establishment of the similarity metric and is part of how the target and model are conceptualized.

With a conceptualization of the target and model into properties in hand, the scientist can add elements to Δ. For example, an ecologist might include terms like "equilibrium abundance" and "maximum population size." For Schelling's model, relevant terms might include "segregated clusters of size *n*," "racial exposure index *r*," spatial layouts of cities and neighborhoods, and descriptions of various movement rules and utility functions.

As science progresses and more is known about a model's targets, the contents of Δ may change. Modelers might initially deem some elements of models and targets important, but these might fall by the wayside as the science progresses. Similarly, new properties of targets might come to be recognized as especially important. These changes in practice and interest will occasion a change in Δ and, consequently, a reevaluation of the model/world relationship. These changes alone can have the effect of rendering the model more or less similar to a target. At first, this might seem like a disadvantage of the account, suggesting that the account's flexibility precludes it giving a good analysis of the model/world relationship.

However, there are two reasons why this is not a disadvantage. First, the theory of similarity I am developing supervenes on properties of the model, properties of the target, and the modeler's intentions. When context or scientific goals change, these intentions will change, and aspects of the relation will change. Second, there are cases in which the perceived quality of a model changes, without any new information about a model or a target. This can happen when a better model is created, but the old one continues to have heuristic value.

5. Modeling Goals and Term Coefficients. While the elements of Δ need to be specified with respect to specific models, targets, and contexts, a more general account can be given about the coefficients for the terms in equation (5). In order to do this, let me begin by addressing an ambiguity in my dis-

cussion of the model/world relation thus far. It is traditional to say that the model/world relation is the relationship in virtue of which studying a model can tell us something about the nature of a target system. But at the same time, scientists are often interested in comparing the relationship that the model actually holds to the world to the one that they are interested in achieving between the model and the world.

In isomorphism-based accounts of this relationship, the only guidance that can be given is that the model is isomorphic or it is not. There is no way of expressing the existence of a relatively good fit between model and target or the gradual improvement of this relationship with improvements to the model. In contrast, the account I am developing allows scientists to asses how close they have come to meeting their goals. It also recognizes that such goals can require different kinds of similarity relations, or at least the emphasis of different kinds of features. This is accounted for by the way in which coefficients for each term of equation (5) are set.

The simplest case is what we can call hyperaccurate modeling. In this type of modeling, the theorist wants the model to contain all of the features of the target and to have neither distortions ($M - T$) nor approximations and further abstractions ($T - M$). In this case, the theorist aims for

$$\frac{|M_a \cap T_a| + |M_m \cap T_m|}{|M_a \cap T_a| + |M_m \cap T_m| + |M_a - T_a| + |M_m - T_m| + |T_a - M_a| + |T_m - M_m|} \rightarrow 1. \tag{6}$$

An advantage of my account is that this need not always be the case. To take just one example, in how-possibly modeling, the goal is to find some mechanism or other that can reproduce the attributes of the target. This means that the attributes of the model and target must be similar, but any plausible mechanism can be used to generate these attributes. The theorist wants to show that some plausible mechanism can produce the phenomenon under investigation. This corresponds to $|M_a \cap T_a|$ having a high value and $|M_a - T_a|$ having a low value. We can express the goal of how-possibly modeling as

$$\frac{|M_a \cap T_a|}{|M_a \cap T_a| + |M_a - T_a|} \rightarrow 1. \tag{7}$$

Other modeling goals might be to represent a core causal factor, to maximize the accuracy of the model's predictions, or to learn about a complex system using multiple models (Weisberg 2007). Each of these kinds of modeling could be represented with different coefficients attached to equation (5)'s terms.

6. Weighting Function and Background Theory. The final aspect of the weighted-feature-matching relation is the weighting function $f(\cdot)$. In very general terms, the purpose of this function is to tell us the relative importance of elements and combinations of elements in Δ. While all features of models and targets that are included in the feature set are taken to have some importance in establishing similarity between model and target, some are considerably more important. The weighting function tells us the relative importance of each feature.

We can build up the general form of the weighting function by considering its general properties. In order to satisfy equation (5), we need to define $f(\cdot)$ over $\mathcal{P}(\Delta_a) \cup \mathcal{P}(\Delta_m)$. However, it is very unlikely that scientists could have anything remotely resembling a representation of a function being defined on this set, or even produce such a function if called on to do so. Moreover, it is unlikely that the relative importance of features is thought of or would be articulated in this way. Rather, scientists typically think about the relative weights of some or all of the elements of Δ. This means that we can substantially restrict the weighting function by requiring that the total weight given on some set X will be equivalent to adding up the weight given to all of the elements of X. This would require the theorist to have access only to the weight she places on each element of Δ.

This seems more realistic but still far from how most scientists think about the model/world relation. Even this restricted form of the weighting function assumes that scientists represent the weight of each element in Δ. But in most cases, scientists will believe that some subset of the features in Δ are especially important and might have a relative weighting of these features. We can call this subset the set of *special features*, and these will be weighted more heavily than the rest. The others will simply be equally weighted. As a default, the weighting function will return the cardinality of sets like $M_m \cap T_m$. But for the subset of features that are special, these will be assigned weight greater than one.

Restricting the weighting function in this way raises a new question: How do scientists determine which elements of Δ are the special features? And for those features, what weights should be put on them? In the best case scenario, background theory will tell us about which terms require the greatest weights. But in many cases of modeling in biology and the social sciences, background theory will not be rich enough to make these determinations. In such cases, the basis for choosing and weighting special features is less clear. What happens in such cases?

In cases with weakly developed background theory, the possibility of reasonable disagreement increases. There can and will be reasonable disagreements about which terms should be weighted more heavily in these cases. However, there is a sense in which choosing a weighting function is in part

an empirical question, and the appropriateness of any particular function is determined using means-ends reasoning.

Much of the time, this procedure is implicit and becomes part of what Kitcher calls a community's *practice* (1993). When assumptions about aims and the relative importance of different aspects of models are widely shared, details about weighting functions are rarely articulated. In fact, in such cases, there may be a range of permissible weighting functions accepted by the community. But when anomalies accumulate, or different subcommunities regard models very differently, communities are forced to be more explicit about their weighting functions. Being explicit in this way can help scientists negotiate their differences.

7. Conclusion. We are now in a position to better articulate the connection between highly idealized models, such as Schelling's model of segregation, and real-world phenomena, such as segregation patterns in Philadelphia. The relationship between the model and the real-world target should be understood as one of similarity. Although highly idealized, Schelling's model would be informative if it shared important features with the real population dynamics of Philadelphia and did not neglect too many important features.

More generally, we can say that models are similar to their targets when they share many, and do not fail to share too many, features that are thought to be salient by the scientific community. This notion of similarity begins from an everyday notion but rejects the idea that similarity is a strictly holistic relation of resemblance. The additional structure of weights and feature sets lets us capture the similarity judgments made by scientists, who may know all the same empirical facts but judge the similarity of a model to its target differently.

REFERENCES

Attneave, F. 1950. "Dimensions of Similarity." *American Journal of Psychology* 63:516–56.
Cartwright, N. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.
Giere, R. N. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.
Goodman, N. 1972. "Seven Strictures on Similarity." In *Problems and Projects*. Indianapolis: Bobbs-Merrill.
Kitcher, P. 1993. *The Advancment of Science.* Oxford: Oxford University Press.
Muldoon, R., T. Smith, and M. Weisberg. 2012. "Segregation That No One Seeks." *Philosophy of Science* 79:38–62.
Quine, W. 1969. "Natural Kinds." In *Ontological Relativity and Other Essays*. New York: Columbia University Press.
Schelling, T. C. 1978. *Micromotives and Macrobehavior.* New York: Norton.
Shepard, R. N. 1980. "Multidimensional-Scaling, Tree-Fitting, and Clustering." *Science* 210:390–98.
———. 1987. "Toward a Universal Law of Generalization for Psychological Science." *Science* 237:1317–23.
Tversky, A. 1977. "Features of Similarity." *Psychological Review* 84:327–52.
Tversky, A., and I. Gati. 1978. "Studies of Similarity." In *Cognition and Categorization*, ed. E. Rosch and B. Lloyd. Hillsdale, NJ: Erlbaum.
Weisberg, M. 2007. "Three Kinds of Idealization." *Journal of Philosophy* 104 (12): 639–59.