# Building More Robust System Dynamics Models Through Validation

William Schoenberg, University of Bergen, Norway
and Jeremy Swartz, University of Oregon, United States

Today human and nonhuman societies face many complex problems that require understanding, not merely through individual relationships, but entire systems so as to recognize unintended consequences. This paper presents an overview of system dynamics, from modeling to critiques, as well as an explanation of where the structural origins of behaviors arise. The discussion includes white box vs. black box models, what makes models useful, and how to improve the quality of models via structural validation, as well as how to make use of the power and adaptability of machine learning. To improve the quality of models, we argue that feedback loop dominance profiles assist to illuminate the underlying causal structure, thus clarifying the feedback-based explanation of dynamics. Methods of feedback system neural networks and feedforward artificial neural networks are offered as ways to do exploratory data analyses which can help to ease model conceptualization processes. From industrial, municipal, global dynamics, and their limits, the fields of system dynamics and machine learning offer new and emerging insights moving pragmatist inquiry forward in the 21st century.

## *Modeling, Systems, and Dynamics*

Modeling as a concept has existed since the dawn of science. At its core, science seeks to understand nature and the universe. The underlying language of science is mathematics, which codifies science's understanding of the universe in the forms of axioms, causal relationships, and natural laws. The organization of related sets of mathematically codified understandings is the very definition of a *model.* One of the greatest stories of modeling in human history is astronomy, where humans have since the dawn of antiquity collectively built a *gigantic* model to explain the motions, and origins of the Earth, the Sun, the stars, galaxies and the universe. This model is astrophysics – all of its laws, theories, and axioms combined to create a mathematical model which can be used to understand, e.g., the retrograde motion of Mars, or the movement of galaxies through the giant voids of the universe. Via that quest to understand all the objects in the universe, science has constantly developed and redeveloped entirely new models of physics and chemistry, starting from the classical physics of the ancient Greeks, Indians, and Chinese, to the scientific revolution and the development of Newtonian physics, or even modern-day quantum and particle physics. All of these developments are the evolution of a model built from a/the mathematically codified understanding of the world around us. The very core of science, the scientific method, is constructed around the development and testing of models, and only those models that prove useful in generating understanding of the real-world phenomena that they represent survive.

Colloquially, systems primarily are the natural phenomena modeled, although the word has a wider meaning when applied in the context of models themselves. For instance, going back to astronomy, our solar system is just that – a *system*, a set of processes, which act on our sun, the planets and their moons, as well as all the asteroids, planetoids, gases and other matter contained within our celestial neighborhood. That (natural) system, a set of processes, is modeled using a system of equations built from the laws of physics and is called celestial mechanics. But the

colloquial language we've so far developed is not fully specific between the natural world and our human representations and understandings of it. Therefore, by taking a wider, more holistic view, the meaning of the word system is a set of interrelated processes that come together to produce dynamics. The individual behaviors which celestial mechanics describe in the context of our solar system include the spin of the Earth around its axis, or the movement of Mars through space as it orbits the Sun. The instantaneous and individual behaviors of the bodies within the solar system which happen in each moment, looked at together over a continuum of time and across a set of interacting bodies, are celestial *dynamics*. Dynamics do not occur in isolation of time or by single elements alone; dynamics are the outcomes of a system composed of at least two elements observed over time. Dynamics apply broadly across systems, both natural and mathematical. As long as there is interaction between the elements producing behavior over time, there are dynamics. All of the work that science does to model systems is centered around creating an understanding of *dynamics* or the cause and effect interactions which exist between things.

The current treatment of modeling used thus far has been exceedingly broad in its scope, but the underlying practices and disciplines are relatively easy to categorize. This paper focuses on mathematical models, specifically mathematical (differential equation) models, as opposed to physical models. Within the domain of mathematical modeling there are two general high-level categories of models: white box and black box. *White box models* are the oldest and most widespread and all of the examples referred to so far are white box models based on explicit systems of equations, where the meaning of each variable and of those equations which house them, have direct and interpretable meaning. In white box models each variable (or node) within the network of model equations that represent a system have a known and agreed upon real-world analogues from the natural system being modeled.

On the other hand, *black box models* are a relatively more modern tool with their origins traced back to the 1950s with the advent of artificial neurons, and later on artificial neural networks, random forests, etc. Black box models are directly and fully based on experimental data, where the understanding they encode in the form of their mathematical structure is not directly interpretable. Each of the nodes within their network of model equations does not have a direct known analogue to the natural system they represent. To be clear, both white and black box models have an underlying mathematical structure. The key difference is the encoding of the understanding of the natural system which those models represent within their mathematical structure. White box models present that understanding in the form of interpretable causal relationships, whereas black box models present that understanding via more opaque mathematical structures, such as networks of artificial neurons which do not provide any specific meaning when studied element by element.

Within the realm of white box models, which are the primary focus of this paper, there are further sub-categories, which are useful for understanding the breadth of modeling techniques. There are models constructed of systems of differential equations (analytical or simulated), or models constructed by the interactions of agents (actors, particles, etc.) and models which create dynamics based upon the generation of a discrete sequence of events, among various others. The ways to construct, interpret and analyze white box models are nearly as boundless as science itself, but that doesn't mean there aren't some useful abstractions that we can apply across wide

swaths of those white and even black box models to help better integrate them all into our understanding.

The first of these useful abstractions across all systems and models are *states*, which pertain to the memory of a system (natural, or mathematical) and provide the system with continuity across time as it creates dynamics. In a model, states are designed to represent the same quantities they do in their natural system in an attempt to mimic that the natural system the model purports to represent. In models, states have direct analogues to their natural equivalents. Examples of states are the number of people in a city, or jelly beans in a jar, or the location of a planet in space. This doesn't mean that states have to be physical quantities, but can be used to track soft concepts, like the amount of knowledge that a person or group of people have about an idea, etc.

A state represents a fundamental element of a system whose value at any given moment in time is dependent upon its previous value and any new changes brought about by the other elements within the system between the current moment and the next. For instance, if time within a system were to stop, the states are the parts of the system that would still be measurable and have a value. For instance, consider a basic system of a bathtub filling with water where there is a spigot which releases water to fill the tub, and a drain at the bottom of the tub which empties it. The amount of water in the bathtub is a state of that system; other states would be the size of the spigot or the size of the drain. Again, both exist outside of time and can be measured without the passage of time. If we wanted to know how much water would be in the tub in the next instant, we would have to know how much water was in the tub in this instant, how much water is added to the tub between this instant and the next, and how much water leaves the tub through the drain or evaporation, etc. during the same time interval. The amount of water which flows into the tub via the spigot or leaves via the drain are the rates which change the value of the states over time. Therefore, rate is the key property which defines a state. States are the fundamental elements of systems whose value can only be changed over time. The water in the bathtub cannot rise or lower in a single instant without the passage of time. Water must be added over some interval of time, no matter how infinitesimal that interval is.

The second useful abstraction across systems and models is *feedback*, or an abstraction used to represent processes which self-modify their own states over time without continued action from a source of action. Let's take, for example, a simple population model of planet Earth. In this overly simplistic model, there is one state which represents the number of people, and two rates which modify the value of that state over time. First, the birth rate which represents new arrivals added to the population, and second, the death rate, which represents existing members of the population who leave through death. In this simple model, there are two feedback processes. The first is the more people in the population, the more births there are, which means over time (not withstanding anything else), there will be more people. This is a *positive* (or *reinforcing*) feedback loop which leads to, if left unchecked, *exponential* behaviors of either growth or decay. In other words, dramatic changes to the situation. The second feedback process in this system says that the more people there are in a population, the more people die, which means fewer people in the population. This is a *negative* (or *balancing*) feedback loop which leads to *logarithmic* stabilizing behaviors.

Now that we have this set of abstractions for looking at and understanding systems and models, how can we apply that to the different popular modern day white box modeling paradigms e.g., agent-based, discrete event, and system dynamics models? First, agent-based models use states to capture the attributes of actors (individuals) or items according to their location in space, their color, etc. The interactions between agents can be looked at through the lens of feedback – any action that an agent takes which modifies one of its states that later comes back to affect that same state of that same agent, is a feedback loop. Finally, the actions of agents or the impacts of other agents on an agent of interest are often delayed in time. Discrete event models can be looked at the same way, as they are structures composed of states linked by feedback whose rates rather than being computed in a continuous fashion, over a series of uniform time intervals, change their states over a discrete continuum of time or in non-uniform intervals of time. Lastly, system dynamics models are composed of *stocks* (or *states*) and *flows* (or *rates*), in other words, feedback loops which connect the stocks and often create time delays. Each of these abstractions can be recognized in the network of model equations generated by all of these modeling paradigms.

In the context of black box models (e.g. machine learning, specifically artificial neural networks)[1], states can be thought of as nodes. In recurrent neural networks, where nodes' values are dependent upon their previous states, it would be fair to call each recurrent node a state. In supervised and reinforcement learning feedback exists in the context of training, but not in the mathematical structure of the final model itself. Training connects the feedforward relationships from the input nodes to the output nodes using weights and biases set during backpropagation that depend upon the outputs of the neural network to re-adjust those weights and biases. This feedback loop between outputs, and weights and biases back to outputs does not exist in the final form of the network of equations of a standard multilayer perceptron artificial neural networks. In fact, for the very large majority of artificial neural networks no feedback at all is present in the final structure.

Now that we've laid the groundwork for understanding systems and models, we turn to the problem of defining and measuring what makes a model useful. The context of this discussion will be on white box, differential equation models, e.g., models built according to the system dynamics methodology. It is often noted that "All models are wrong, but some are useful." This begs the questions: what makes a model useful? how can we measure a model's utility? and how can we be confident that the ways in which a model are wrong doesn't affect the learning we are trying to do using the model?

## *System Dynamics and its Modeling Process*

System dynamics[2] as a field seeks to understand via simulation models[3] the underlying *nonlinear*[4] behavior of systems using accumulations (states), feedback and time delays and grew out of Jay W. Forrester's research in the 1950s at the Massachusetts Institute of Technology (MIT). In the early 1960s, system dynamics was focused primarily on modeling and understanding problems in business or management as Forrester was a professor at the recently formed Sloan School of Management at MIT and published *Industrial Dynamics* in 1961. In the late 1960s, Forrester's *Urban Dynamics* (1969) demonstrated the beginnings of the broad utility of the system dynamics methodology with the study of the management of a city. Branching out

from businesses and cities ultimately led him to the Club of Rome in 1970, where he became deeply interested in global socioeconomic modeling and managing global systems, publishing *World Dynamics* in 1971. Subsequently, *World Dynamics* laid the groundwork for *The Limits to Growth* (Meadows, et al., 1972), which became a seminal publication because it was one of the first to take the position that humanity could grow to surpass the physical carrying capacity of its habitat, planet Earth. (Also see *The 30-Year Update*, Meadows, et al., 2004).

Since the 1970s, there has been a continuing history of good modeling work coming out of the system dynamics field including Peter M. Senge's *The Fifth Discipline: The Art & Practice of the Learning Organization* ([1990] 2006), John D. Sterman's *Business Dynamics: Systems Thinking and Modeling for a Complex World* (2000), and more recent work such as HealthBound, PRISM, ReThink Health, and all the work done by Climate Interactive. As a method, system dynamics is very broad and generally applicable to a wide variety of systems, deriving power and utility from its laser focus on the understanding of systems regardless of the context of those systems. In a general sense, the purpose of the system dynamics method is to help people make better decisions in the face of complexity (Sterman, 2000).

After reading the above you may think that system dynamics is a magical method with the ability to shed light and understanding on a huge variety of systems across a wide domain of subjects. To better understand the critical discourse around system dynamics, it is important to understand the system dynamics method as it is practiced and implemented.

In 1994, Forrester described the practice of *system dynamics* in general as the process of understanding and improving systems. In the same paper, Forrester laid out the general steps of the modeling process shown in Figure 1. The process is clearly not linear, and exposes many opportunities for iteration and revision.

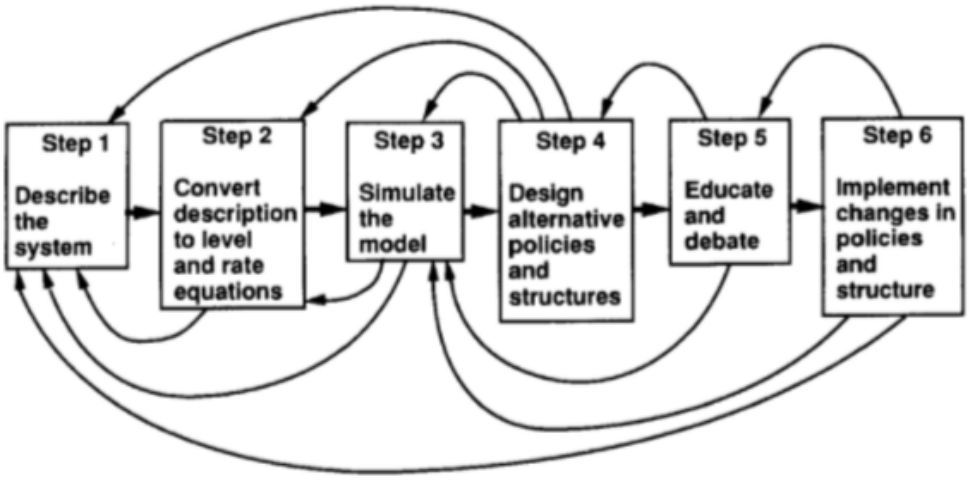


FIGURE 1: The system dynamics modeling process (Forrester 1994).

Figure 1 illustrates how the system dynamics process at its core generalizes real-world observations into complex (computational) mathematical models, which can be seen in steps one and two. It then uses those mathematical models as a basis for developing, then sharing inferences (policies) which are then used as the basis for either education, and/or making

decisions in the real-world, which can be seen in steps three through six. As an example, we can apply this process to understanding a typical business issue of declining revenues, or an environmental problem such as climate change. Step 1 would be to describe how the business operates, or how products are demanded, produced, and sold. For the climate system it would be to enumerate how the Earth's various natural systems work together to regulate the climate. The second step would be to encode that understanding into a series of equations (a model), making sure to get more descriptions of the system as questions about the specifics of the problem arise. Step 3 involves the modeler simulating that model (which can require that more equations be written, or more information be solicited). If done properly, this results in a depiction of the original problem by the model. In the fourth step, the modeler looks for solutions to the problem in the model. For the business example of declining revenue, perhaps by adding more salespeople, lowering the price, etc., which may influence each of the previous steps. For the climate system by adjusting human GHG emissions, or land use rates. Step 5 is to use the model to educate the key stakeholders about what the model says and why, which includes the possibility of learning more and therefore adjusting previous steps. Thus, in the sixth step, the stakeholders can implement the change and either fix the original problem which would end the process, or go back to any of the other steps and understand other ways to improve the model).

Delving deeper into some of the philosophy of science issues involved at the foundation of the field of system dynamics requires an understanding of empiricism and rationalism as they apply to the field. Yaman Barlas and Stanley Carpenter (1990) developed and refined these philosophies into worldviews as they apply to system dynamics. The large majority of the critical sentiment surrounding system dynamics related to the validation of models and determining model utility. Barlas and Carpenter's work was summarized by Barlas (1996), who classified the philosophies of science surrounding system dynamics into two opposing philosophies. The first is the traditional reductionist/logical positivist worldview that takes a theoretically valid model as an objective representation of a real system. Under this worldview, the model is either right or wrong and by matching it with real-world empirical data, the model's usefulness is determined. The second worldview is the relativistic, holistic pragmatism that takes a valid model as just one of many possible understandings of a real-world system, and views a model as a vessel for its author's worldview. To this second perspective, models aren't right or wrong in a binary state, but exist on a sliding scale somewhere between the two concepts. Barlas and Carpenter (1990) argued that the second worldview is most complementary to the system dynamics approach.

Early criticisms of the work of Forrester and other system dynamists is typified by the neoclassical economist William D. Nordhaus, in his paper, "World Dynamics: Measurement Without Data" (1973), which was written in response to Forrester's *World Dynamics* (1971). Nordhaus took issue with Forrester's neo-Malthusian position and criticized the method of system dynamics and computer simulation by saying "…without an accurate model there is no assurance that systems dynamics is better than mental models; the main result is a spurious and misleading precision." (p. 1157) He argued that the model was "measurement without data" and that none of the relationships between the variables in the model are couched in empirical studies or observed data. Essentially Nordhaus was claiming that by abstracting to such a degree as Forrester did, all meaning and connection with the real-world was lost and that inferences drawn

from this model could not in any way contribute to the understanding of that real-world system, again a traditional reductionist/logical positivist perspective.

Nordhaus continued by pointing out that there was insufficient validation done through empirical testing. Applying a reductionist/local positivist perspective to this work would result in an argument that insufficient sensitivity analysis was performed on this model yielding an incorrect perception of accuracy and precision. In fact, one of the things Nordhaus did in his paper was to perform a sensitivity analysis on the parameters and relationships in the world dynamics model to demonstrate the possible range of outcomes and policies that could be produced. As a key proponent of the second holistic, pragmatist worldview, I believe that Forrester would argue against Nordhaus by saying that the model has value because it informs discussion and thinking on the topics of boundless economic and population growth.

More recently, the system dynamics process been expanded with a few steps added to the high-level process shown in Figure 1. Validation and uncertainty testing now are done on any model before it can be realistically used to provide policy recommendation or serve as a basis for education in response to the criticisms of the subscribers to the traditional reductionist/logical positivist worldview. Barlas described the process for model validation in his seminal work, "Formal Aspects of Model Validity and Validation in System Dynamics" (1996). Barlas' major departure from the process is to place emphasis on the modeler to ensure that their models are structurally and behaviorally valid between simulation and policy analysis. At a very high-level, behavioral validation means that the model is generally capable of reproducing historical data with a high degree of accuracy without the model having foreknowledge of that history. And at the same high-level, structural validation refers to the concept that the model is producing that behavior for the right reasons, e.g. that the system of equations that the model is composed of are logically and empirically valid.

Building upon the previous examples of a business with declining revenues or global climate change, between the simulation and policy steps (3 and 4), as well as between the policy construction steps and the education steps (4 and 5), the modeler now must explicitly focus on validation, where it was more implicit before. For instance, in the business example the modeler must check that their model reproduces the decline in revenue with a high-level of accuracy. In the climate case, they must confirm that their model accurately portrays historical climate data. They may confirm with subject matter experts and existing bodies of literature that the system of equations used to encode the knowledge that they gathered about the system is logical and borne out by data. Other tests include assessing their model's sensitivity to parameters to make sure point estimates are not the primary cause for behavior, or to identify specific assumptions which must be further tested and validated. They typically also subject their model to extreme conditions tests to make sure that, for instance in the business case, when there are no salespeople, there are no sales, etc. At its core, Barlas' addition is to make sure that the modeling process emphasizes the scientific process of exposing a theory to data to try and find cases where the model doesn't explain the data.

Barlas produced a list of all the validation tests a model must go through before being considered useful for policy recommendations. He breaks these tests down into three categories: direct structure tests, structure-oriented behavior tests, and behavior pattern tests. He demonstrates

how these tests confirm and validate structure (the relationships between the variables that Nordhaus criticized Forrester for), as well as behavior and also quite importantly cover model purpose, and problem identification (the issue of boundaries).

*Addressing the Criticisms of Modeling (System Dynamics)*
*via Automated Feedback Loop Dominance Analysis*

System dynamics models are constructed through the heavy use of feedback loops. A typical system dynamics model of consequence to policy can contain as few as 3-5 feedback loops or as many as tens of millions, as was the case for the model underlying *Urban Dynamics*. By studying the relationships between the feedback loops in a system dynamics model, we can understand where model behavior comes from, and via the process of induction, relate that back to real-world processes that are responsible for the problem being modeled.

The normal mode of operation for model understanding today comes from years of experience and training through either (and oftentimes both) repeated practice or training during a Masters and/or Doctoral education. In the literature, this process is often referred to as the art of modeling or model analysis. Richardson (1996) perfectly captures the essence of this process:

> For more than 35 years practitioners have relied on a time-consuming and often incomplete process that iterates from formulation to parametrization, testing, observation, hypothesizing and back again. […] Understanding connections between complex model structure and behavior comes, if one is skillful and/or lucky, after a prolonged series of model tests of deepening sophistication and insight.

This process requires much experience, knowledge and practice and sets a high barrier to entry into the field of useful system dynamics modeling in a policy making context, which acts in part as a safe-guard for the policy maker who must ultimately undertake any responsibility for making actionable decisions based upon the recommendations from system dynamics models. On the other hand, this process is error prone and may not yield a proper understanding of model behavior and therefore may lead to real-world determinants of system behavior which can drive policy makers to make sub-optimal decisions based on improperly understood stories and narratives of system behavior.

From the perspective of a traditional reductionist/logical positivist, the art of the model analysis approach described above is a major problem. There is no clear right and wrong and understanding is purely in the eye of the beholder. The traditionalist worldview within the system dynamics field has developed two rarely used (in a policy context), and hard to apply (even from an expert perspective) mathematical techniques for objectively performing this process of model understanding.

The current methods (both mathematical and art of analysis) relate model behavior back to the underlying feedback loop structure. This process whether done by hand or algorithm is called feedback loop dominance analysis. The current state of the art in the system dynamics field for performing feedback loop dominance analysis relies on either practitioner intuition and

experience (the art of modeling and model analysis) or complex algorithmic feedback loop dominance analysis. The former is taught as part of the methodology of model building, while the latter comes from 40 years of work on techniques to derive and explain model behavior based on the analysis of structure (for example, see Graham, 1977; Forrester 1982; Eberlein, 1984; Davidsen, 1991; Mojtahedzadeh, 1996; Ford, 1999; Saleh, 2002; Mojtahedzadeh et al., 2004; Goncalves, 2009; Saleh et al., 2010; Kampmann, 2012; Hayward and Boswell, 2014; Moxnes and Davidsen, 2016; Oliva, 2016; Sato, 2016; Hayward and Roach, 2017; Naumov and Oliva, 2018; Oliva, 2020).

David Ford (1999, pp. 4-5) most clearly states the needs of the system dynamics field as it relates to loop dominance analysis:

> To rigorously analyze loop dominance in all but small and simple models and effectively apply analysis results, system dynamicists need at least two things: (1) automated analysis tools applicable to models with many loops and (2) a clear and unambiguous understanding of loop dominance and how it impacts system behavior.

Loop dominance analysis sheds light on the origins of behavior in system dynamics models by relating observed behavior back to the feedback process(es) that created it (Forrester 1961; Richardson 1991). Loop dominance analysis is concerned with the discovery of the strength and polarity of the key feedback loops existent within models as time progresses within those models (Richardson, 1995). Over any period of time in a simulating model, some feedback loops are most important to the expressed dynamics of the model above all others, and those feedback loops are referred to as the dominant feedback loops. Loop dominance is a concept measured at each and every specific point in time when the model simulates. This means during some points in time (for instance, in a simple SIR [Susceptible-Infectious-Recovered] epidemic model), the behavior of the model early on may be dominated by a reinforcing feedback loop which spreads the diseases to new susceptible suitable hosts, but later on the behavior of the model may come to be dominated by a balancing feedback loop which limits the spread of the disease due to running out of susceptible, suitable hosts. This continuum or pattern of shifting feedback loop dominance is called the feedback loop dominance profile.

By identifying the feedback loop dominance profile of a model, we better understand the causes for the observed behavior over that period of time from a mathematical perspective. For instance, by changing the gain of the feedback loops by changing model parameters or inputs, or even by modifying the loop structures themselves, we modify the feedback loop dominance profile of the model, and of course change the behavior of the model. In other words, the explanation of the causes for the dynamics exhibited by the model changes as the structure or inputs into that model change. However, it must be noted that the feedback loop dominance profile on its own says nothing about whether or not the model is structurally valid and relevant to the problem being studied. It still requires expert assessment to validate if the measured feedback loop dominance profile captures the essence of the underlying system being modeled. To a traditionalist, the ability to identify the feedback loop dominance profile of a model (assuming it is a perfect representation of the underlying system) is the key to the development of robust policy options, to perform impact assessment by way of simulation and to the formulation of policy

recommendations. This is because by understanding which feedback processes are most meaningful and how they contribute to the dynamics being changed (again only if you assume your model is a perfect representation of reality), demonstrates complete mastery of the system and the ability to intervene safely and productively, knowing everything that will happen is a consequence of your intervention.

From the perspective of the traditional reductionist/logical positivist once you move past the issues typified by Nordhaus and addressed by Barlas' emphasis on model validation (both behavioral and structural) you're left with just a single strong criticism of the system dynamics approach, and generally speaking, of modeling as a general practice. The core of that criticism is that modelers cannot provably demonstrate that their models are actually representative of the systems they purport to embody. Validation is a process of elimination, subjecting the model to batteries of tests aimed at finding flaws, where the model clearly and importantly misrepresents reality. The validation techniques in use today both within the system dynamics field, and modeling in general do not fully address the problem that a model may not be truly representative of the system it purports to understand. Each of the validation practices in use today—confirming that models are able to replicate historical dynamics without foreknowledge, or confirming that the equations within a model are each logical and valid in isolation—helps to build confidence that the model is an accurate representation of the system. The paradigm here, and of science in general, is that you can never prove a model right only wrong. Within that paradigm there is another step that we can take to address the concern that a model may not be representative of the system it purports to embody via feedback loop dominance analysis. One of its holy grail pursuits within the field of system dynamics is a mathematical system capable of performing loop dominance analysis algorithmically using a mathematical process to understand which feedback processes are responsible for the behavior of the model at each and every point in time (Sterman, 2000).

Recently there has been a break though in automating model analysis via an easy to use, and widely applicable automated method for objectively measuring feedback loop dominance called *Loops That Matter* (LTM) depicted in Figure 2 (Schoenberg et al., 2019). LTM tells the modeler which feedback loops are responsible for producing the dynamics generated by the model at each point in time while it simulates. It does this by analyzing each equation in the model to determine the importance of each independent variable to its dependent variable(s) and uses that information to generate a measurement of feedback loop importance which is compared across all feedback loops in a model (Schoenberg et.al, 2019, Schoenberg et. al, 2021). By knowing the importance of each link and feedback loop in the model is, modelers can speak with confidence about where specifically in their model behavior is being generated. The theory underlying LTM is very broad in its scope, even while the implementation of the algorithm is currently limited in its application to system dynamics style models. The primary reason that LTM represents a major breakthrough is because it can be used as a very powerful new structural validation test measuring whether or not the model produces behavior for the right reasons, or in other words, that the feedback loop dominance of the model matches with an empirically derived understanding of the system which the model represents. For instance, we could rule out the SIR epidemic model as useful if during the exponential spread of the disease, the model's behavior is being dominated by a feedback loop or set of loops which has nothing to do with disease spread. By algorithmically calculating the feedback loop dominance profile, LTM makes it abundantly

clear why a model produces the behavior it does, which allows the modeler and other subject matter experts to validate whether the model provided reasonable explanations of the underlying system.

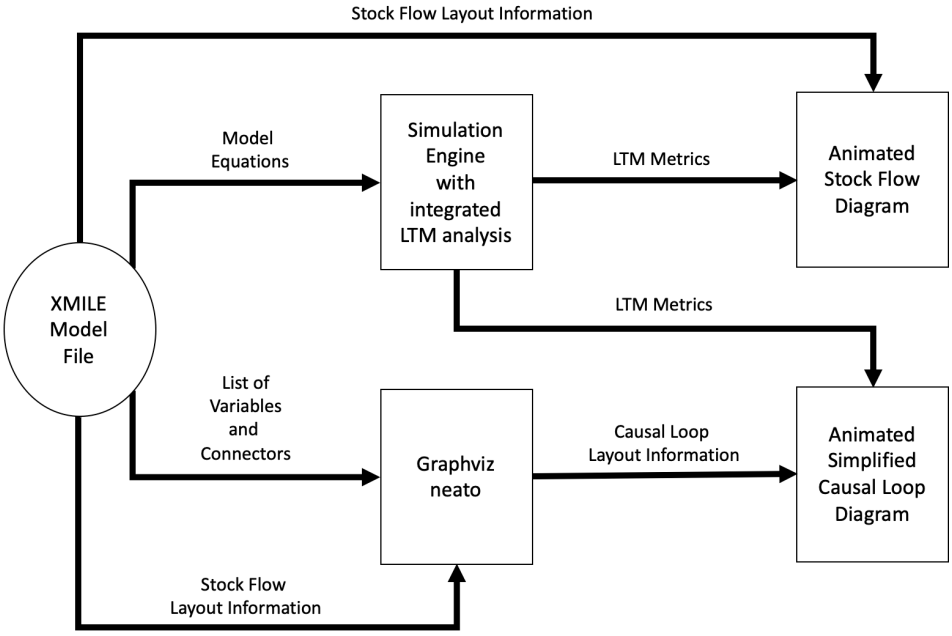

FIGURE 2: Diagram of how Loops That Matter (LTM) process works.

Therefore, to address the aforementioned philosophical criticism of system dynamics; that models may not properly represent reality, the system dynamics modeling process needs to be once again amended. The change adds an LTM analysis to the model validation step so that the generated explanation for dynamics, the feedback loop dominance profile, can be validated against both the pre-existing understanding of the system being modeled, and the empirical data gathered about that system. Thus, modelers can be assured that if their model passes the existing battery of validation tests, it is much less likely to be random coincidence. Therefore, the model is a much more useful tool for answering questions where we do not understand how the system will respond to perturbations. So even though "all models are wrong," we'll at least know which models are useful, and that those will be the models in which the descriptions of where behavior comes from match reality in all cases where we currently understand the system under study.

### *Turning a Black Box Model into a White Box Model*

Modern day *machine learning*[5] methods including probabilistic modeling, kernel machines or deep learning, have arisen from a strong, almost single minded, focus on empiricism making use of the extraordinary amounts of observational data available from a plethora of sources (Ghahramani, 2015; Schölkopf and Smola, 2008; Goodfellow, et al., 2016). When viewed through the lens of accurate prediction power, machine learning has proven to be quite successful, but present-day techniques typically fail to reveal the fundamental causal mechanisms driving behavior. Although it cannot be ignored that interpreting the structure behind black box, deep learning models is an active research area (Montavon et. al, 2018). To make full use of the technological advancements in machine learning, significant emphasis must

be placed on finding a valid and interpretable causal understanding of the underlying real-world system (Runge, et al., 2019).

The study of observational causal inference exists alongside the field of machine learning and is focused on drawing conclusions about causal connections between variables by studying the response in an effect variable when a cause is being changed (Pearl, 2009). Largely based on statistics, observational causal inference started with the seminal works of Norbert Wiener and Clive W. J. Granger and has been growing and developing since the 1950s (Wiener, 1956; Granger, 1969). The most well-known method is *Granger causality*[6], which tests whether omitting the past of a time series $X$ in a time series model including $Y$'s own and other covariates' past, increases the prediction error of the next time step of $Y$ (Granger, 1969). Granger causality is useful in discovering specific causal links in a system, but it fails to generalize to complex nonlinear systems, typically failing to identify all of the links in the networks of feedback relationships which govern these systems from a holistic perspective (Spirtes and Zhang, 2016).

Non-Granger methods in observational causal inference have been categorized into the following three broad frameworks: nonlinear state-space methods, causal network learning algorithms, and structural causal models (Runge et al., 2019). These approaches are designed to discover causality in directed acyclic graphs and are incapable of operating in complex dynamic feedback-rich environments, which has been argued to be at the core of the most intractable and relevant problems (Sterman, 2000). Taking into account the current state of the art in machine learning and automated observational causal inference, it becomes clear that the machine learning field lacks a direct understanding of the relevance of the causal structures underlying their models to real-world systems.

The power of the LTM method for discovering the origins of behavior in models has direct utility outside of the field of system dynamics. By combining states, feedback loops, and feedforward artificial neural networks, we're now able to do automated causal inference using a machine learning approach and relying on LTM to turn the generated black box model into a white box model in all the ways that matter, namely, the ability to generate a feedback loop dominance profile (Schoenberg, 2019). This approach, called a *feedback system neural network* (FSNN), constructs a model with one state for each dimension in the dataset. Each state is connected to each and every other state (including itself) using a standard *multilayer perceptron artificial neural network* (MLPANN) forming a system of ordinary differential equations, a system dynamics model with a factorial of the number of states (stocks) feedback loops, typically an amazingly large number (trillions, quintillions, a Googleplex or more!). That model is parameterized (trained) via calibration (back-propagation) to fit the empirical data. That parameterized black box model is then analyzed using LTM to discover the origins of behavior via its feedback loop dominance profile ultimately turning the black box model transparent enough to reason about, visualize and understand what the model has learned about the data, and therefore to validate the generated model's mathematical structure.

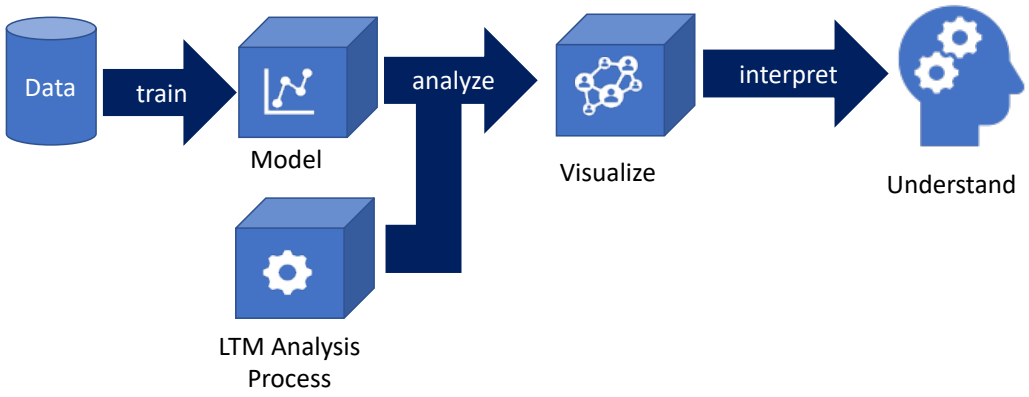


FIGURE 3: High-level overview the machine learning approach to causal inference including LTM.

The FSNN method (presented in Figure 3) acts as a machine that produces behaviorally accurate, feedback rich structural hypotheses directly from data where the polarity and contribution of each link between all states in the system is known. The machine takes time series data measured from the real-world as its input, and, with no additional input, produces a behaviorally relevant causal dynamic hypothesis, which is much more easily validated by subject matter experts than standard neural nets and other machine learning technologies. The validation performed also is of a much higher quality because it is a structural validation of the machine learned understanding of the system rather than a behavioral validation of the outputs of the model.

By combining system dynamics and machine learning FSNN's demonstrate a powerful new machine learning technique for finding the causality in arbitrary data sets. By using MLPANNs to represent the relationships between the states the universal approximation theorem (Cybenko, 1989) ensures that each relationship is capable of reproducing the real-world function which links the states together. By linking all states together in a directed cyclic graph this new machine learning method becomes much more likely to find accurate causal models for the data. As Jay W. Forrester might say, this is because nonlinear feedback systems are at the heart of complex dynamic real-world problems and those systems are best represented in mathematics as systems of ordinary differential equations containing feedback, which this new method does as a matter of purpose.

FSNNs are not cure-alls which replace the existing system dynamics modeling process. They are most useful when you're trying to describe and understand the potential causal relationships in the system in the early phases of the modeling process. FSNN's are useful for exploratory data analysis, but are not meant to be the end of the process, but rather at the beginning. They are especially useful for intractable problems which exist in inscrutable systems where there is a lot of data, but precious little understanding of the high-level relationships which govern the processes responsible for generating the observed dynamics. An example of such a case may be in genomics where there is a glut of empirical evidence on gene expression, but much less understanding of how those gene expressions are related.

*Using Modeling to Solve Problems and General Conclusions*

Today human and nonhuman communities face many complex problems, including environmental change, food insecurity, worsening inequality, etc. Solving these problems requires understanding, not merely individual relationships, but entire systems, so as to recognize unintended consequences. The tools, techniques, and problem-solving, or in other words, *technologies* (Swartz, et al., 2019), necessary to make progress on the development of society may be based in part on models, whether white box or black box. Following the processes laid out by Forrester and amended by Barlas is an excellent start to developing models, but as previously suggested, there are at least two steps forward that modelers and systems scientists can embrace.

First is to improve the quality of models by doing better structural validation using LTM to produce a feedback loop dominance profile that illuminates the underlying causal structure, clarifying the feedback-based explanation of dynamics. Whether the model is system dynamics, agent-based, or black/white box, it is no longer sufficient to state that a model represents a system because it reproduces behavior or has some pieces of logical structure which can be verified as existing in the natural system. Modelers must now demonstrate that their models reproduce the behavior of the systems they purport to represent for the right reasons, by demonstrating the feedback loop dominance profile of their models matches what is known about the systems in reality. Without this, those models are objectively less useful because there is no assurance that they will accurately respond to non-historically observed inputs. Essentially, the link between model and reality is severed!

Second, modelers can use techniques like FSNNs to make use of the amazing power and adaptability of machine learning models to quickly build hypotheses to do exploratory data analysis which ought to ease the model conceptualization process. The major overarching emphasis here is on structural validation and ensuring that models are useful because they demonstrably capture the essence of reality for more reasons than just behavior matching in the past.

The solutions to the aforementioned major problems facing the world are not generated by technologies (including analytics) alone. A key to developing robust societies capable of tackling systemic issues is *systems education*. Organizations like the Creative Learning Exchange[7] and others who work to get systems thinking and systems sciences integrated into the K-12 education curriculum are vital to the future abilities of societies to solve these challenges and potentially avoid future intractable problems. It's not just the scientists, or the modelers, or the analysts who need to be trained to think systemically. It is also the educators, policy makers, stakeholders, artists, and everyone living, breathing and functioning as parts of systems—all of whom can bring forth interventions to solve and validate difficult problems. From industrial, municipal, global dynamics, and their limits, the fields of system dynamics and machine learning offer new and emerging insights into public health, epidemiology, systems biology, environmental science, restoration ecology, public policy, disaster management, education, and beyond.

REFERENCES

Barlas, Yaman (1996), 'Formal aspects of model validity and validation in system dynamics', *System Dynamics Review,* 12(3), pp. 183-210.
Barlas, Yaman, and Stanley Carpenter (1990), 'Philosophical roots of model validation: Two paradigms', *System Dynamics Review,* 6(2), pp. 148-166.
Cybenko, George (1989), 'Approximation by superpositions of a sigmoidal function', *Mathematics of Control, Signals and Systems*, *2*(4), pp. 303-314.
Davidsen, Pål I. (1991), *The Structure-Behavior Graph*. The System Dynamics Group, Cambridge, MA: MIT Press.
Eberlein, Robert Larry (1984), 'Simplifying dynamic models by retaining selected behavior modes', Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Ford, David N. (1999), 'A behavioral approach to feedback loop dominance analysis', *System Dynamics Review,* 15(1), pp. 3–36.
Forrester, Jay W. (1961), *Industrial Dynamics,* Cambridge, MA: MIT Press.
——— (1969), *Urban Dynamics,* Cambridge, MA: MIT Press.
——— (1971), *World Dynamics,* Cambridge, MA: Wright-Allen Press.
——— (1994), 'System dynamics, systems thinking, and soft OR', *System Dynamics Review*, 10(2-3), pp. 245-256.
Forrester, Nathan Blair (1982), 'A dynamic synthesis of basic macroeconomic theory: Implications for stabilization policy analysis', Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Ghahramani, Zoubin (2015), 'Probabilistic machine learning and artificial intelligence', *Nature*, *521*(7553), p. 452.
Gonçalves, Paulo (2009), 'Behavior modes, pathways and overall trajectories: Eigenvector and eigenvalue analysis of dynamic systems', *System Dynamics Review,* 25(1), pp. 35–62.
Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016), *Deep Learning*, Cambridge, MA: MIT Press.
Graham, Alan Karl (1977), 'Principles of the relationship between structure and behavior of dynamic systems', Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Granger, Clive W. J. (1969), 'Investigating causal relations by econometric models and cross-spectral methods', *Econometrica: Journal of the Econometric Society*, pp. 424-438.
Hayward, John, and Graeme P. Boswell (2014), 'Model behaviour and the concept of loop impact: A practical method', *System Dynamics Review* 30(1), pp. 29–57.
Hayward, John, and Paul A. Roach (2017), 'Newton's laws as an interpretive framework in system dynamics', *System Dynamics Review,* 33(3-4), pp. 183-218.
Kampmann, Christian Erik (2012), 'Feedback loop gains and system behaviour', *System Dynamics Review,* 28(4), pp. 370–395. [Originally published in *Proceedings of the 1996 International System Dynamics Conference*, Cambridge, MA: Systems Dynamics Society, pp. 21-25.]
Meadows, Donella H., Meadows, Dennis L., Randers, Jørgen, and Behrens, William W., III (1972), *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*, New York: Universe Books.
Meadows, Donella H., Randers, Jørrgen, and Meadows, Dennis L. (2004), *Limits to Growth: The 30-Year Update*, White River Junction, VT: Chelsea Green Publishing.
Mojtahedzadeh, Mohammad T. (1996), *Structural Analysis of the URBAN1 Model*, University at Albany, State University of New York (SUNY).

Mojtahedzadeh, Mohammad, David Andersen, and George P. Richardson (2004), 'Using Digest to implement the pathway participation method for detecting influential system structure', *System Dynamics Review*, 20(1), pp. 1-20.
Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller (2018), 'Methods for interpreting and understanding deep neural networks', *Digital Signal Processing*, *73*, pp. 1-15.
Moxnes, Erling, and Pål I. Davidsen (2016), 'Intuitive understanding of steady-state and transient behaviors', *System Dynamics Review*, 32(2), pp. 128-153.
Naumov, S. and Oliva, Rogelio (2019), *Structural Dominance Analysis Toolset*, System Dynamics Group, Massachusetts Institute of Technology, Cambridge, MA.
Nordhaus, William D. (1973), 'World dynamics: Measurement without data', *The Economic Journal* 83.332, pp. 1156-1183.
Oliva, Rogelio (2016), 'Structural dominance analysis of large and stochastic models', *System Dynamics Review,* 32(1), pp. 26-51.
——— (2020), 'On structural dominance analysis', *System Dynamics Review*, 36(1), pp. 8-28.
Pearl, Judea (2009), 'Causal inference in statistics: An overview', *Statistics Surveys*, *3*, pp. 96-146.
Richardson, George P. (1991), *Feedback Thought in Social Science and Systems Theory*, University of Pennsylvania, Philadelphia, PA.
——— (1995), 'Loop polarity, loop dominance, and the concept of dominant polarity', *System Dynamics Review*,11(1), pp. 67–88.
——— (1996), 'Problems for the future of system dynamics', *System Dynamics Review,* 12(2), pp. 141-157.
Runge, Jakob, Bathiany, Sebastian, Bollt, Erik, Camps-Valls, Gustau, Coumou, Dim, Deyle, Ethan, Glymour, Clark, Kretschmer, Marlene, Mahecha, Miguel D., Muñoz-Marí, Jordi, van Nes, Egbert H., Peters, Jonas, Quax, Rick, Reichstein, Markus, Scheffer, Marten, Schölkopf, Bernhard, Spirtes, Peter, Sugihara, George, Sun, Jie, Zhang, Kun and Zscheischler, Jakob (2019), 'Inferring causation from time series in Earth system sciences', *Nature Communications*, *10*(1), p. 2553.
Saleh, Mohamed M. (2002), 'The characterization of model behavior and its causal foundation', Ph.D. dissertation, University of Bergen, Bergen, Norway.
Saleh, Mohamed, Oliva, Rogelio, Kampmann, Christian Erik, Davidsen, Pål (2010), 'A comprehensive analytical approach for policy analysis of system dynamics models', *European Journal of Operational Research*, Volume 203(3), pp. 673-683.
Sato, Jeremy B. (2016), 'State space analysis of dominant structures in dynamic social systems', Ph.D. dissertation, Washington University, St. Louis, MO.
Schoenberg, William ([2019] 2020), 'Feedback system neural networks for inferring causality in directed cyclic graphs', arXiv preprint arXiv: 1908.10336. [2019 (version 1), 2020 (version 2)]. https://arxiv.org/abs/1908.10336. Accessed 24 March 2021.
Schoenberg, William, Pål Davidsen, and Robert Eberlein (2020), 'Understanding model behavior using the loops that matter method', *System Dynamics Review,* 36(2), pp. 158-190.
Schoenberg, William, Hayward, John, and Eberlein, Robert (2021), 'Improving loops that matter', paper presented at the 2021 System Dynamics Conference.
Schölkopf, Bernhard, and Alexander J. Smola (2008), *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond,* Cambridge, MA: MIT Press.

Senge, Peter M. ([1990] 2006), *The Fifth Discipline: The Art & Practice of the Learning Organization*, Revised and Updated Edition, New York: Doubleday.

Spirtes, Peter, and Kun Zhang (2016), 'Causal discovery and inference: Concepts and recent methodological advances', *Applied Informatics*, Vol. 3, No. 1, p. 3.

Sterman, John (2000), *Business Dynamics: Systems Thinking and Modeling for a Complex World*, New York: Irwin/McGraw-Hill.

Swartz, Jeremy, Wasko, Janet, Marvin, Carolyn, Logan, Robert K. and Coleman, Beth (2019), 'Philosophy of Technology: Who Is in the Saddle?', *Journalism & Mass Communication Quarterly*, 96(2), pp. 351-366.

Wiener, Norbert (1956), 'The Theory of Prediction', in E. F. Beckenbach (ed.), *Modern Mathematics for the Engineer: First Series,* New York: McGraw-Hill, pp. 165-190.

---

[1] *Machine learning* algorithms are typically broken down into 3 high level categories – supervised learning, unsupervised learning, and reinforcement learning.

- *Supervised learning* is when the input data to the algorithm contains paired sets of inputs with outputs, and the algorithm develops a model using that input data which can map any arbitrary input to its specific and correct output.
- *Unsupervised learning* is when the input data to the algorithm does not contain any outputs. An unsupervised learning algorithm creates a model which classifies and groups the given data as it sees fit without any preconceived notion for how the inputs should be arranged into outputs.
- *Reinforcement learning* is when an algorithmically generated model receives feedback about its output as it navigates the problem space. Rather than pairing each input with a specific output (supervised learning), the algorithm trains the model by giving it rewards as it produces certain outputs.

*Deep learning* refers specifically to artificial neural networks with multiple layers where a deep neural network is one which allows for universal approximation as defined by the universal approximation theorem.

[2] ßThe history of the tools used to develop system dynamics models may be of interest to some readers. The earliest system dynamics (SD) modeling language was DYNAMO developed in the late 1950s at MIT and has a programming language like interface. The first of the visual SD tools based around programmable stock and flow diagrams was STELLA developed by High Performance Systems (now isee systems) and was first released in 1985 for the Macintosh. Other popular visual SD modeling tools are Vensim (produced by Ventana), and Powersim which was initially called Constructor, now Studio. Other tools exist, these four though are the most notable.

[3] As opposed to systems of differential equations which are analytically solved which are typically referred to as Dynamic Systems.

[4] Nonlinear behavior is when a change in the inputs to the system do not create a proportional change in the outputs.

[5] Machine learning refers to computer models (algorithms) which self-modify their state through trial and error (optimization, backpropagation) on a given set of empirical data.

[6] Granger causality is a statistical concept which uses prediction to measure if a change in an independent variable causes a change in a dependent variable. The technique is based on linear regression, although there are nonlinear methods which are difficult to apply.

[7] The Creative Learning Exchange, http://www.clexchange.org/cle/about.asp.