Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, L. H. Lee, P. Lendermann, eds.

PRINCIPLES AND REQUIREMENTS FOR SIMULATION-DRIVEN INCREMENTAL LEARNING OF CAUSAL EXPLANATORY MODELS

Levent Yilmaz

Department of Computer Science and Software Engineering
Auburn University
3116 Shelby Center
Auburn, AL 36849, USA

ABSTRACT

In Agent-based models of complex adaptive systems, emergent behavior is not engineered but results from local interactions among agents. Due to the consequences of complex, distributed interactions among decentralized agents, the causal chain of cross-cutting processes that give rise to emergent behavior is difficult to discern and explain. To provide a context for the explainability of Agent-based models, a systematic review of philosophical and cognitive models of causal explanation is provided. For illustration purposes, the theory of explanatory coherence is used as a computational framework for learning explanatory cognitive maps of increasingly refined and broadened model features. The framework offers a perspective that signifies principles for learning causal explanatory models with implications for simulation model development environments.

1 INTRODUCTION

Computational models are often complex monolithic entities that are difficult to comprehend and explain (Davis et al. 2018). Agent-Based Models (ABMs) that rely on autonomous agency, adaptation, and context-sensitive interactions are particularly challenging to understand due to emergent behavior (Egli et al. 2018). This challenge is akin to the explainability crisis in Artificial Intelligence that calls for interpretable systems that can communicate their reasoning (Gunning, David 2017). Similar concerns exist in the broader modeling and simulation domain, partly due to the significance of reproducibility and transparency in the use of models (Yilmaz 2012; Teran-Somohano et al. 2014), as well as the need for establishing trust in simulations (Onggo et al. 2019; Yilmaz and Liu 2020).

Because cause-effect reasoning is a critical objective in science, the concept of explanation has long been a central focus of model-based science (Magnani 2009; Bokulich 2017). The goal of scientific inquiry is to discover an explanation Y for a phenomenon in terms of the underlying mechanisms that generate it. Similarly, engineers search for a design mechanism, X, which can produce a desirable property, Y, for an envisioned system. In both cases, scientists and engineers are driven by an intrinsic motivation to understand or generate targeted aspects of natural or artificial systems. To support achieving this objective, models have become instrumental in scientific thinking and communication of explanations (Gelfert 2019; Magnani 2009). The theory and methodology of modeling can benefit from insights offered by the philosophy, psychology, and cognitive science of science to advance its role further. Such insights can help discern what abstractions are conducive to effective explanations for (1) conveying causal relations to others, including learners, and (2) facilitating programmable abductive model building to generate plausible alternative explanations.

Besides, explanations can be used to study models independent of a target system. That is, one can explore a model's behavior across its entire range to explain when and which tentative features cohere together for generating expected behavioral regularities. Analytical support for causal reasoning about model behavior has a wide range of applications. For instance, explanations are central to sensemaking by facilitating the understanding of events and the formation of causal mental models. They help model users determine what would happen in counterfactual reasoning or intervene to support prospective anticipatory thinking by projecting into future states.

In practice, the provision of explanatory support via causal models improves understanding and helps formulate better questions and adapt a model to better align with the objectives of a simulation study. This paper reviews and analyzes the extant literature on the philosophy and psychology of explanation. The overview lends itself to a methodical strategy grounded on the theory of explanatory coherence and reflective equilibrium for generating explanations of ABMs and outlining plausible strategies for learning explanatory models through simulation experiments. The requirements and principles of explanatory modeling are delineated to provide a foundation for next-generation simulation infrastructures that support the development of self-aware and explanatory models.

2 BACKGROUND

Both science and engineering involve model-based discovery and explanation activities. Therefore, issues concerning explanation have long been central to the philosophy of science, aiming to characterize normative models and criteria for causal reasoning. Normative models often use idealized cognitive strategies such as logic-based deduction, inductive generalization, and abductive reasoning. On the other hand, because explanation involves cognitive effort and needs to relate to mental models of the intended audience, cognitive models of explanation need to be considered as well. Such an understanding is necessary to develop pragmatic model-based explanations. In this section, we examine philosophical, cognitive, and theoretical foundations of model-based explanations, in which the explanations refer to properties and behavior of the idealized abstract model in describing the observed regularities.

Explanations that explicitly relate causes and effects by appealing to causal claims aim to address why or how something has occurred. Such explanations can either be causal-mechanical, process-centric representations that focus on the organization of entities and activities or rely on construals based on theory and data. One of the earliest models of explanation is the Deductive-Nomological (DN) model (Hempel and Oppenheim 1948), which views an explanation as a deductive argument. Probabilistic and inductive methods extend the DN model for statistical explanation to support explanation under uncertainty. Deductive-Statistical explanations involve deriving regularities based on more general statistical laws, whereas Inductive-Statistical explanations attach a likelihood to outcomes based on the probabilistic interpretation of explanans. The Statistical Relevance (SR) model (Salmon 1971) leverages conditional dependence relations among events to provide explanations that deductive or inductive strategies cannot capture. Specifically, given attributes, X, Y, and Z, the attribute Z will be statistically relevant to attribute Y if $P(Y|X \cap Z) \neq P(Y|X)$; that is if the probability of event Y conditional on X and Z is not the same as the probability of Y conditional X.

Inspired by the DN and AR models, the *simulacrum* account of explanation (Cartwright and McMullin 1984) requires finding a model that fits into the basic framework of theory and serves as an analog of the system. Similarly, the theory of explanatory coherence (Thagard 1989) provides a sound computational framework to establish relations of local coherence between conjectured explanations, including propositions about regularities in observed data. The inference to the best explanation takes the form of measuring how well the conjectured explanation coheres with observations and other accepted principles and explanations.

The abductive reasoning process starts with a trigger. In ABM, this may be an emergent behavior that is either surprising, and thus requires explanation, or expected and hence needs an explanatory justification. According to (Peirce 1992), this observed event needs to be nontrivial. Following the observation of the event, one or more explanations are generated. This process is often a creative act, but context, prior domain knowledge, and heuristics help narrow relevant explanations. Generated explanations are then evaluated to identify possible more likely explanations to support the observed event. The *manipulative abduction* strategy underlines the role that the context and its creative manipulation play to reveal new explanatory hypotheses (Magnani 2009).

Consistent with the AR model, (Thagard 2012) views scientific explanation from a cognitive perspective and categorizes explanation into three major processes and four critical technical methods. The processes are (1) selecting an explanation based on available information, (2) creating new hypotheses for an explanation, and (3) evaluating plausible competing explanations to make an inference to the best explanation. The methods commonly used as part of these processes are then categorized into deductive (e.g., rule or logic-based strategies), schematic models based on explanation patterns and analogies, probabilistic (e.g., Bayesian models), and connectionist network models of explanation.

When there are multiple explanations, the selection process is not always based on utility and expected value calculations. For instance, in the theory of explanatory coherence (Thagard 1989), explanations are evaluated based on coherence judgments via constraint satisfaction mechanisms. The selected tentative explanation is expected to be reevaluated for revision and refinement based on further evidence. The use of coherence as a strategy for explanation is also evident in the unificationist account (Friedman 1974) of explanation. A theory that unifies a broad range of phenomena provides a compelling account of explanatory relevance. Under the unificationist view, understanding something fits it into a broader pattern. The wider the pattern, the more its explanatory power. Pattern-oriented explanations are also analogical due to their ability to explain distinct phenomena that can be subsumed by the same set of schematic explanation patterns.

The Causal Mechanical (CM) model of explanation (Salmon 1984) is a process-centric theory of explanation. Because CM emphasizes the significance of tracing spatio-temporal processes in formulating explanations, it has gained traction in the context of discovering and explaining biological mechanisms (Craver and Darden 2013). The causal mechanistic view considers entities, their activities, and their organization as the central elements of an explanation. The activities within and between entities produce, underlie, and maintain the regularities observed in the system (Darden 2002). In accord with the CM model, an explanation is characterized as information that is relevant to manipulation and control, implying the synergy between explanation and exploration (Woodward 2005). The manipulation of the causes facilitates *counterfactual reasoning*, which determines the variation in the explanandum had the factors referenced in the explanans been different.

3 EXPLANATORY COHERENCE MAPS

Agent-based models are often used to specify complex adaptive systems comprised of a large number of spatially connected entities that interact with each other and the context, resulting in emergent behavior that is difficult to attribute to specific features of the model. Designing ABMs that explain emergent behavior observed in a system is a *target-directed explanation* activity. By finding a model capable of generating the targeted regularity, scientists conjecture a plausible explanation. Following the convergence to a plausible model, *model-directed explanation* shifts the focus to exploring the consequences of the model's assumptions and explaining how the consequences manifest as a function of the features of the model.

3.1 Theory of Explanatory Coherence

In seeking a balance between alternative explanations, beliefs, and judgments, a strategy is needed to attain a state of coherent justification that accounts for the explanation. According to the *Reflective Equilibrium* theory (Daniels 1996), in the presence of conflicts among beliefs, judgments, goals, and explanations, humans proceed by adjusting beliefs until they are in equilibrium. In the most general sense, Reflective Equilibrium can be construed as an attractor state in a complex adaptive system.

The attractor state emerges at the end of a perceptual and deliberation process by which we reflect on and revise our explanations and goals about an area of inquiry. If we can view equilibrium as a stable state that brings conflicts to a level of resolution, the equilibrium state serves as a coherence account of justification. An optimal equilibrium can be attained when there is no further inclination to revise judgments because together they have the highest degree of acceptability (Daniels 1996; Thagard 2002). The principles and judgments that one arrives at when the equilibrium is reached can account for the context, and the situation examined. The *theory of explanatory coherence* provides a set of principles that lends itself to a computational strategy in the form of constraint satisfaction to compute the state of reflective equilibrium (Thagard 2002).

The constraint satisfaction strategy is similar to viewing a state in an N-dimensional space. The activation levels of explanatory hypotheses or nodes are analogous to the acceptability of the respective propositions. Each node receives input from and is reinforced by every other node that is explanatorily connected. The inputs can then be moderated by the weights of the link from which the input arrives. The activation value is updated as a function of the weighted sum of the inputs it receives. The process continues until the activation values of all units settle. Formally, if we define the activation level of each node j as a_j, where a_j ranges from -1 (rejected) to $+1$ (accepted), the update function can be defined as follows:

$$a_j(t+1) = \begin{cases} a_j(t)(1-\theta) + net_j(M - a_j(t)), & \text{if } net_j > 0 \\ a_j(t)(1-\theta) + net_j(a_j(t) - m), & \text{otherwise} \end{cases}$$

In this formulation, the variable θ is a decay parameter that decrements the activation level of each unit at every cycle. In the absence of input from other units, the activation level of the unit gradually decays, with m being the minimum activation, M denoting the maximum activation, and net_j representing the net input to a unit, as defined by the following equation: $\sum_i w_{ij} a_i(t)$. These computations can be carried out for every node until the activation levels of elements stabilize and the network reaches an equilibrium via self-organization. Nodes with positive activation levels at the equilibrium state can be distinguished as maximally coherent propositional explanations.

As an illustrative example, consider the Prey-Predator dynamics, which is used in modeling the behavior of biological systems with multiple interacting species. These species play the roles of prey and predator, and their populations change as a result of competitive as well as cooperative interactions.

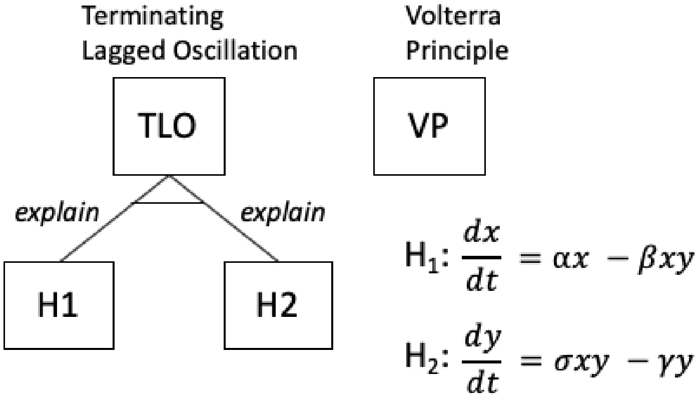
Terminating
Lagged Oscillation
TLO
Volterra
Principle
VP
explain
explain
H1
H2
H_1: $\frac{dx}{dt} = \alpha x - \beta xy$
H_2: $\frac{dy}{dt} = \sigma xy - \gamma y$

Figure 1: Explanatory relations. Lotka-Volterra equations are used to explain population dynamics in the presence of two species. The rate of change of population sizes is described in terms of parameters that describe interaction between species.

Figure 1 depicts two significant properties of interest, lagged oscillation and the Volterra principle, along with two hypothesized equation-based mechanisms that are conjectured to explain these properties.

H1 (prey dynamics) and H2 (predator dynamics) together explain the lagged oscillation behavior while falling short of exhibiting behavior consistent with the Volterra principle. Such equation-based models are highly idealized and easy to explain abstracted phenomena; however, they become intractable as the complexity increases due to factors associated with the context and the specific activities of the individual members of the population. Equation-based models provide averages across populations while ignoring details that involve theories of community structure, measures of environmental diversity, food chain and stability, and diverse individual behavior.

3.2 Abductive Reasoning with Explanatory Cognitive Maps

To improve the accuracy and realism of explanations, increasingly detailed and refined hypotheses are introduced to explain such generic hypotheses and to account for the impact of specific factors such as community structure and individual activities. ABMs mitigate the concerns associated with equation-based models, by describing idealized equation-based models in terms causal mechanisms that involve entities, their activities, and interactions. On the other hand, as ABMs introduce programmatic simplifications and utilities that are not necessarily connected to the underlying theory, determining whether the results are due to essentials of the model or its externalities becomes a challenge. For instance, consider the following model that aims to realize H1 and H2 with two species: prey and predator. The rules for the predator agent are as follows:

- **Movement rule**: Move one step in the random direction.
- **Consume rule**: Check if there is a prey at the current location. If there is prey, randomly select one prey and gain energy by consuming it.
- **Reproduction rule**: If the agent has sufficient energy, the agent picks a random number between 0 and 1. If the number is less than or equal to the predator reproduction rate, it spawns a new predator agent at its current location.
- **Death rule**: If the agent has an energy level of 0, then the predator dies.

Notice that these rules roughly represent the constraints of the equation-based model in the original Lotka-Volterra model. Representing an equation-based model in terms of an ABM requires making explicit assumptions about individual behavior that were either implicit or undefined in the equation-based model. Therefore, there can be multiple ABM realizations that aim to generate targeted behaviors such as stabilized oscillations in populations sizes. Determining which one of many alternative, competing models is a valid representation of the system for the intended purposes of the study emerges as a significant challenge.

For instance, the equation-based model does not make any assumptions about the movement of either the prey or predator. The movement rule in the ABM can vary from completely random moves to more realistic representations that include group behavior and risk averse strategies to avoid regions that are considered risky based on experiential learning. An ABM developer needs to make explicit decisions about such representational issues. To complete the model specification and test its ability to generate stabilized oscillation as an emergent behavior of population dynamics, prey rules can similarly be defined as follows:

- **Movement rule**: Move one step in the random direction.
- **Reproduction rule**: Sample a random number. If the number is less than or equal to prey-reproduction-rate, then reproduce.
- **Death rule**: If the agent is caught by a predator, it dies.

Figure 2 depicts which rules are used as refined explanatory mechanisms for each high-level idealized equation model. The birth, death, and move rules are used to characterize the growth rate in H1. Predators' move and consume behaviors are also associated with H1 to account for the rate at which preys and predators interact, possibly resulting in the death of preys. Similarly, the move, consume, and death rules

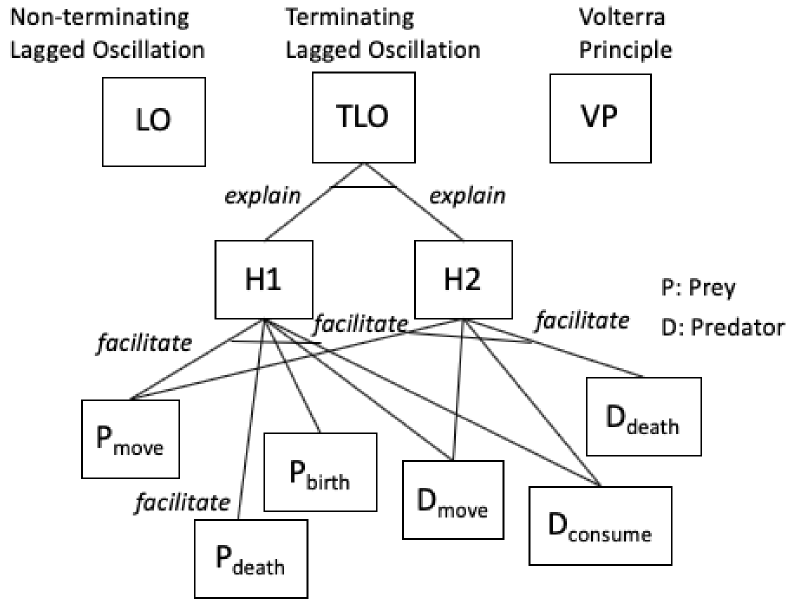

Figure 2: Refinement of explanatory hypotheses. Idealized equation models are replaced with with their respective ABM specifications,

of predators, along with the move behavior of preys are used to develop a causal mechanistic explanation of `H2`. Replacing idealized equation models with their respective ABM specifications, we observe *terminating lagged* oscillation. That is, the model generates lagged oscillation for a period, but the population of either the prey or the predator reaches to 0 or its maximum, resulting in the termination of the simulation. Therefore, the hypothesized behavioral rules are not robust realizations of the original hypotheses. To mitigate this issue, the behavioral rules need to be revised via abductive reasoning by postulating new hypothetical explanations.

Because neither the availability of food nor its density are explicitly modeled, in the original hypothesized equation-based model, the prey population grows exponentially in the absence of interaction with the predators. The size of the spatial context, as well as the density and distribution of either the prey or predator population influence the rate at which they interact. Yet, in reality, the growth of populations are counterbalanced by the degree of availability of food resources. To facilitate such counterbalancing behavior, the reproduction rule of the prey is specified as contingent on the energy gained from the food resource in the environment. To this end, an environment component (i.e., grass) is introduced as a food resource for the prey (i.e., sheep). The density of the food resource is controlled by the `grass-growth-rate`. By consuming the grass, the prey agent receives energy that stimulates the reproduction process. However, uncontrolled increase in the prey population is now suppressed by diminishing levels of food, imposing negative feedback that balances the positive feedback caused by reproduction.

If the environment has resources, the prey's consumption rule is facilitated to help explain the desired lagged oscillation behavior. Because the environment features with and without resource elements are alternatives, they are not compatible and hence contradict and suppress each other. Therefore, they cannot be used as simultaneous explanatory mechanisms. While the environment with the resource feature supports lagged oscillation, the alternative feature explains terminating lagged oscillation, as is the case observed in the previous version of the model.

The cognitive explanatory map shown in Figure 3 suggests which hypotheses and associated behavioral mechanisms can be selected to support a specific targeted behavior. The explanatory power of hypothesized causal explanations increase as they succeed in (1) broadening their support to explain increasing number

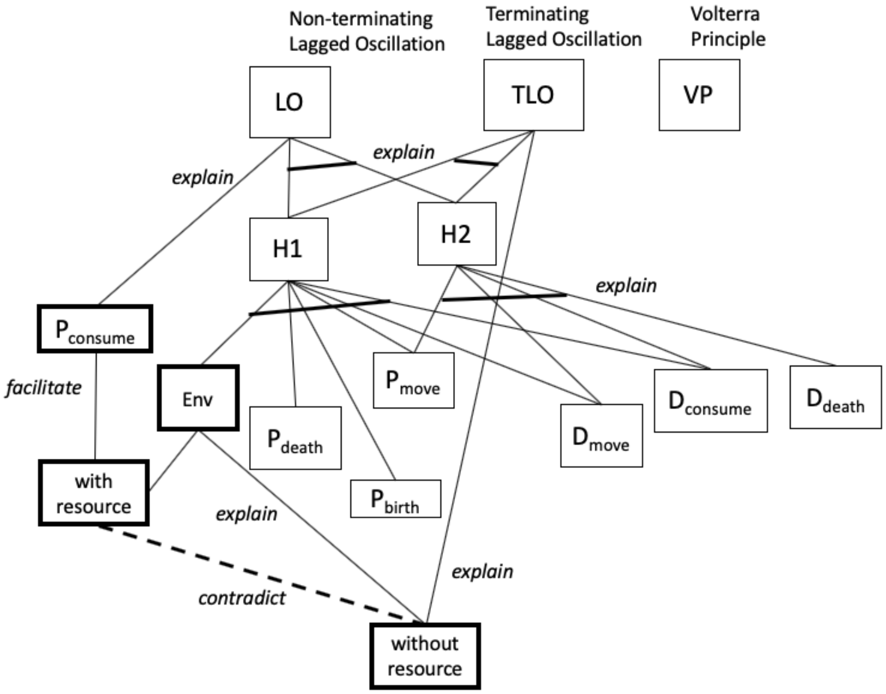

Figure 3: Revising Explanatory Hypotheses. An environment feature with two alternative realizations, with and without resource, is included to further refine hypothesis H1

of targeted behaviors and (2) deepening their explanation in terms of lower-level fundamental mechanisms that explain how and why high-level explanations work. An important finding of the Lotka-Volterra prey-predator dynamics model is the Volterra principle, which is an empirical behavior observed in fishery statistics.

According to this principle, as characterized by H1 and H2, when prey and predator dynamics are negatively coupled, injecting biocide (e.g., toxic element) into the environment increases the abundance of the prey and decreases the abundance of the predator. To broaden the applicability of H1 and H2, the model is modified with a new causal mechanism that introduces *biocide* into the context. Moreover, the consumption behavior of both the prey and predator are refined by two alternative explanatory causal mechanisms; one that allows consuming biocide and the other that considers only the food resource. Figure 4 depicts how H1 and H2, along with the resource consumption and environmental biocide features support explaining the Volterra principle.

The discussion above is intended to highlight the three critical elements of target-directed explanation. First, we need to be able to select a set of coherent explanatory features from among alternatives. However, if existing explanatory features are not sufficient to generate the target behavior, alternative features are created and included in the set of plausible explanations for consideration. Second, explanations are evaluated to make inference to the best explanation. That is, alternative explanatory features are ranked in terms of their degree of relevance and success in generating the desired or expected behavior. Finally, the scope and resolution of explanatory features are expanded to further instill confidence into the explanatory power of identified features. Specifically, the explanations are broadened to target additional behavior while also being refined into increasingly detailed and high-resolution causal mechanistic features.

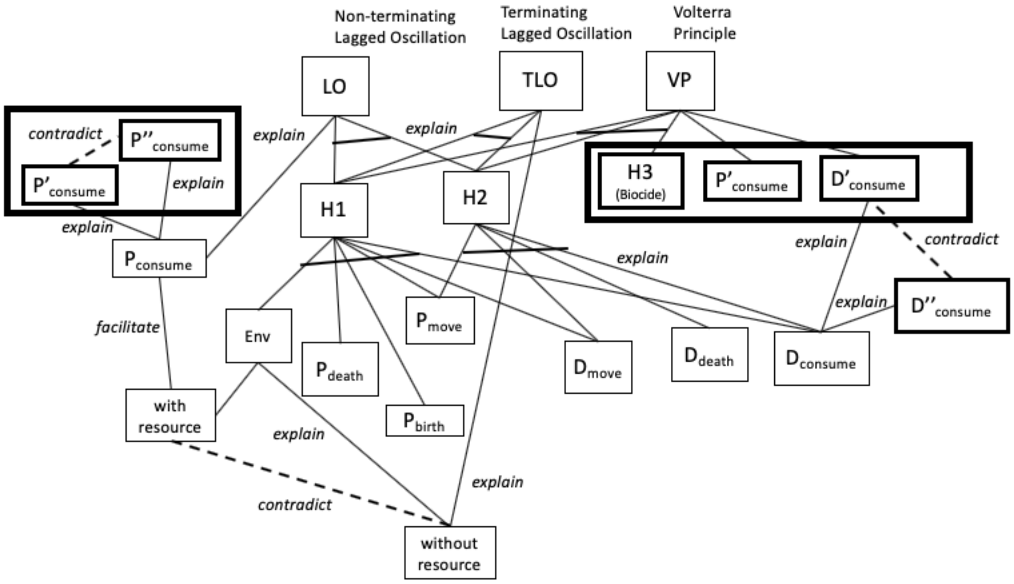

Figure 4: Specific configurations of `H1` and `H2`, coupled with refined resource consumption and environmental biocide features support explaining the Volterra principle.

4 PRINCIPLES FOR EXPLANATORY MODELING

The illustrative example presented above suggests that the generation of explanatory models involves the interplay of multiple activities and is a highly dynamic process that involves specific inferential processes. The example and its underlying coherence-driven strategy motivates the following principles.

Principle 1 – Explanation is an iterative, incremental process: The provision of a model-based explanation stimulates further inquiry to deepen and broaden the scope of the plausible explanations. Initial explanations often provide a template to continue the search process, allowing model builders to iteratively refine the model's causal mechanisms by adding details to increase its level of resolution. During the process, the focus of inquiry can shift due to the evaluation of alternative explanatory mechanisms.

Principle 2 – Explanation is a symbiotic adaptive process: As the process of searching for explanation unfolds, a symbiotic search process takes place between the hypothesis (e.g., model structure) and experiment spaces. Following the search within the architectural space of models, the experimental conditions are created so that they provide new information that would otherwise be unavailable. Such new information is then used to influence the search process within the hypothesis space (Magnani 2009).

Principle 3 – Emergence prompts explanation: Agent-based models of complex systems reveal emergent properties as a result of interactions among a diverse set of agents in a spatial context. Emergent behavior can be unexpected and surprising, as it may be an indicator of new knowledge that cannot simply be inferred as a linear function of agent attributes. The recognition of such emergent behavior prompts the need for explanation to provide an account of causal mechanisms responsible for the observed behavior.

Principle 4 – Explanation requires understanding via self-awareness: Having an introspective capability to assess the consequences as a function of the premises of one's behavior is critical for reflection. Explanatory reasoning requires reflecting on the underlying causes of observed behavior, and such reflection enables the explanation of behavior in terms of beliefs, objectives, and intentions that drive observable actions. A model with such self-awareness capabilities can compare its simulated behavior to objectives and evolve an understanding of its features and how and when they contribute to the desired behavior.

Principle 5 – Explanation involves an exploratory learning process: As shown in Figure 5, the generation of an explanation involves the use of three vital inferential processes: deduction, induction, and abduction. Deduction occurs in the form of simulation of the model to derive the consequences of its underlying assumptions. Initially, there can be multiple competing explanations, each represented by a distinct model. Distinct models in the ensemble represent plausible explanations examined under simulated experiments to reveal their behavior. The exploration phase requires the selection of experiments and the variation of the model to improve the ability of model users to distinguish between alternative models concerning their explanatory power. Simulation-generated data are then generalized via inductive mechanisms to support learning rules that separate successful results from those that fail to satisfy the objectives.

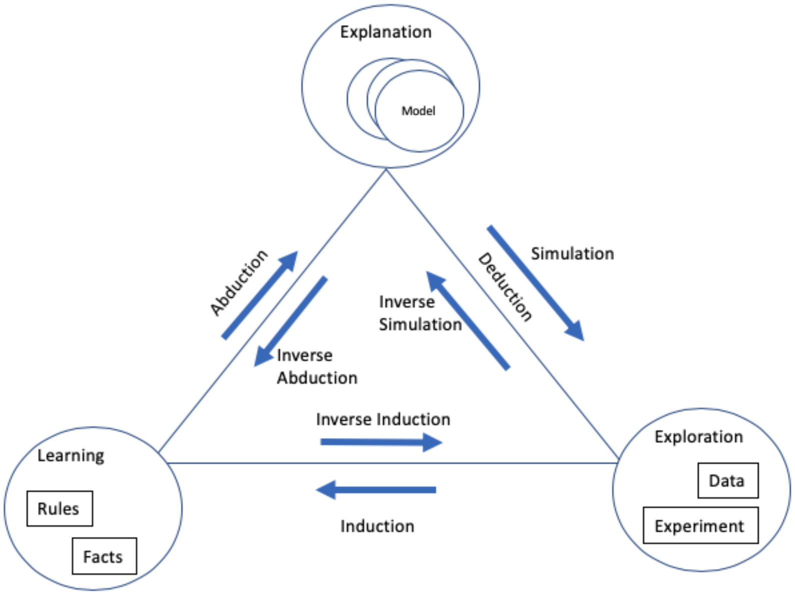

Figure 5: The generation of an explanation involves the use of three key inferential processes: deduction, induction, and abduction.

5 REQUIREMENTS FOR MODELING ENVIRONMENTS

Explanatory modeling involves exploratory learning to support the generation and evaluation of explanations. Exploration requires variations of model structure and representations and the experiments. The ability to generate numerous models with different architectures across multiple resolutions and perspectives must be managed transparently. Such management should be extended to computational experiments so that alternative scenarios and model variants can be coupled adaptively in the context of an evolving analysis.

To generate explanatory models, the exploration results need to be generalized and presented using visual, causal, and conceptual representations meaningful to model users. Furthermore, the set of possible models needs to be explicitly represented to be queried, sampled, and analyzed to support incremental and adaptive modeling. Self-adaptive models impose additional constraints for representing and selecting model features.

5.1 Variation of Causal Assumptions in an Evolving Analysis

Models are developed with specific assumptions in mind. The rules that define agents' actions are often theoretical and pragmatic elements. Pragmatic constructs include simplifying assumptions that are not related to the underlying theory of the system. For instance, the movements of prey and predator agents in an agent-based model of ecological dynamics can be either random or intentional to mimic the flocking behavior of collectives. Supporting seamless variation of model structures and representations can be in implicit model design strategies or based on explicit modeling constructs intended to facilitate systematic variability management. Implicit design strategies include well-known design patterns such as the *Strategy* and the *Factory Method*, which are used in model behavior variation when the specific behavior cannot be anticipated in advance.

Alternatively, explicit modeling constructs and paradigms can be considered to design models with variability in mind. For instance, both the feature and aspect-oriented modeling paradigms promote relatively simple features and cross-cutting aspects that can be assembled into aggregate models subject to composition constraints. Modeling environments can make such constraints and feature models explicit by using declarative specifications, which can allow the use of powerful tools that can manipulate and adapt declarative specification transparently, and hence enable manipulative abductive reasoning (Magnani 2009). A modeling environment needs to support an explicit variability model to support such variation and customization. Such a model can include at least a mechanism for specifying the behavior of individual features, defining acceptable and well-formed feature combinations in terms of syntactic and semantic composition constraints, and providing the means for evolving these specifications as a result of the information gathered through simulation experiments.

5.2 Controlling for Confounding Factors

To facilitate explanations that move beyond the provision of associations, a model development environment needs to facilitate interventions and counterfactual reasoning. Experiments are critical instruments in search of explanations, as they enable intervening with the conditions under which a model is simulated to observe their effects. As such, interventions allow discerning effects of carefully manipulated causes. Whereas interventions focus on prospective reasoning, the counterfactual mode of experimentation involves revisiting the assumptions underlying the model by asking questions such as ”What is the probability of observing the outcome $Y = y$ had we used the causal mechanism `M1`, given that we observed the outcome $Y = y'$ and designed the model with the mechanism `M2`”. While intervention-based exploration requires searching the experiment space, counterfactual analysis requires searching within the model design and experiment spaces.

To experiment on a model is to place it under the control of an experiment manager to prune and revise the space of possible causal mechanisms. The focus on evaluating alternative causal explanations suggests the design of targeted experiments to reveal the role and significance of elements of putative mechanisms. One experiment strategy is to determine whether an entity, activity, property, or organizational feature is causally relevant to another. This strategy helps model developers determine the contextual conditions that trigger a phenomenon or understand if a specific element is sufficient for what happens in the next stage of the mechanism. Another strategy is to discern whether an entity, activity or organizational characteristic is relevant for the overall behavior generated by the whole mechanism. These experiments facilitate determining causal relations across the levels of hierarchically organized models. Once sufficient confidence is achieved, various questions can be answered by intervening with the model and the experimental conditions.

Targeted experiments are mechanism-aware in that they allow direct interference, stimulation, and activation of components in a model. Interference experiments should provide facilities to inhibit, diminish, or disable the components of the underlying causal structure. In *stimulation experiments*, one intervenes to excite or amplify the model's behavior and detect the impact of that change. Interference and stimulation experiments are bottom-up due to their intervention in the lower-level mechanistic components. On the other hand, the activation experiments are top-down experiments that influence the conditions to activate higher-level behavior and detect the impact of such activation on the lower-level components.

Whereas intervention experiments are intended to determine the role of relatively stable mechanisms comprising specific components, the exploratory analysis starts by asking specific mechanistic questions to identify these components in the first place. There are various kinds of such experiments (Darden 2002).

- In *by-what-activity* experiments, the goal is to discover which kind of activity connects the causes to effects. Such experiments require reasoning about the alignment of invariants with the preconditions of activities. One needs to identify the preconditions of plausible activities and compare the situations in which that precondition is met with cases in which it is not. These comparisons can be used to discriminate among competing hypothesized activity features.

- Activities are produced by components that collaborate to produce observed behavior. In *by-what-component* experiments, artificial conditions are created to disable components in a controlled manner to discern whether the intervention prevents the phenomenon. Alternatively, different combinations and organizational representations of components are conducive to generating the desired behavior. Activity-based and component-oriented experiments can be combined to prune the space of possible explanatory causal mechanisms.

5.3 Learning from the Results of Exploratory Simulation

As experiments are conducted to intervene with the context and mechanisms of the target system, the results are expected to gradually converge to increasingly credible mechanisms that are successful in producing and maintaining the desired behavior. As a result, trust in the model's behavior evolves due to its ability to account for a broadening set of accumulating evidence. Successful models continue to be refined to include increasingly detailed characterization of the high-level mechanisms. This learning process should be facilitated through appropriate tools that can evaluate the results of experiments and provide necessary critic to guide the search process further.

Evaluation of the efficacy of the components and activities requires the explicit separation of a learning element from the model so that the *learning component* can use the feedback received from the critic and determine how the model should be modified to perform better concerning the goals of the experiment. The *critic* informs the learning element via reward (or penalty) how well the model is performing for the performance standard. The learning process requires an additional *generative component*, which facilitates model generation in accord with the abductive reasoning strategies outlined earlier. Unless a model generator explores alternative representations, the model can quickly converge to a representation that can perform well under specific scenarios but may fail to behave robustly as the range of scenarios is broadened.

6 CONCLUSIONS

The application domains of successful models broaden while refinements to a model deepen the level of resolution to improve accuracy and fidelity as learning takes place. However, explaining the cause-effect relations in such models is a critical challenge, primarily due to the autonomous, decentralized decision-making by agents that adapt and interact with each other, giving rise to emergent behavior. Explanation strategies are grounded in the philosophy and cognitive science of science are reviewed.

Due to the need for aligning explanations with the cognitive requirements of the target audience, a strategy based on cognitive coherence is adopted. The strategy demonstrates how explanatory models can evolve with a simulation model based on the results of experiments. Specifically, the theory of explanatory coherence and reflective equilibrium can show how an inquiry in population dynamics can be broadened and deepened to revise beliefs across multiple levels and scales about causal premises of expected behavioral regularities. Following the demonstration of the theory of explanatory coherence, the principles of explanatory modeling are delineated to characterize the highly dynamic process that results in the formation and growth of explanatory models. The underlying inferential processes are explained within a framework that interfaces with exploration and learning. Based on the principles and the evaluation criteria, we conclude with specific guidelines for developing modeling and simulation environments that support explanatory modeling.

REFERENCES

Bokulich, A. 2017. "Models and explanation". In *Springer Handbook of Model-Based Science*, 103–118. Springer.

Cartwright, Nancy and McMullin, Ernan 1984. "How the laws of physics lie".

Craver, C. F., and L. Darden. 2013. *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.

Daniels, N. 1996. *Justice and justification: Reflective equilibrium in theory and practice*, Volume 22. Cambridge Univ Press.

Darden, L. 2002. "Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining". *Philosophy of Science* 69(S3):S354–S365.

Davis, P. K., A. O'Mahony, T. R. Gulden, O. A. Osoba, and K. Sieck. 2018. *Priority challenges for social and behavioral research and its modeling*. RAND Corporation Santa Monica, CA.

Egli, L., H. Weise, V. Radchuk, R. Seppelt, and V. Grimm. 2018, aug. "Exploring resilience with agent-based models: State of the art, knowledge gaps and recommendations for coping with multidimensionality". *Ecological Complexity*.

Friedman, M. 1974. "Explanation and scientific understanding". *The Journal of Philosophy* 71(1):5–19.

Gelfert, A. 2019. "Assessing the Credibility of Conceptual Models". In *Computer Simulation Validation*, 249–269. Springer.

Gunning, David 2017, May. "Explainable artificial intelligence (xai)". Defense Advanced Research Projects Agency (DARPA). https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

Hempel, C. G., and P. Oppenheim. 1948. "Studies in the Logic of Explanation". *Philosophy of science* 15(2):135–175.

Magnani, L. 2009. *Abductive cognition: The epistemological and eco-cognitive dimensions of hypothetical reasoning*, Volume 3. Springer Science & Business Media.

Onggo, S., L. Yilmaz, F. Klugl, T. Terana, and C. Macal, M. 2019. "Credible Agent-based Simulation – An Illusion or Only a Step Away?". In *Proceedings of the Winter Simulation Conference*, in–press. ACM.

Peirce, C. S. 1992. *The essential Peirce: selected philosophical writings*, Volume 2. Indiana University Press.

Salmon, W. C. 1971. *Statistical explanation and statistical relevance*, Volume 69. University of Pittsburgh Pre.

Salmon, W. C. 1984. *Scientific explanation and the causal structure of the world*. Princeton University Press.

Teran-Somohano, A., O. Dayıbas, L. Yilmaz, and A. Smith. 2014. "Toward a model-driven engineering framework for reproducible simulation experiment lifecycle management". In *Proceedings of the Winter Simulation Conference 2014*, 2726–2737. IEEE.

Thagard, P. 1989. "Explanatory coherence". *Behavioral and brain sciences* 12(3):435–467.

Thagard, P. 2002. *Coherence in thought and action*. MIT press.

Thagard, P. 2012. *The cognitive science of science: Explanation, discovery, and conceptual change*. Mit Press.

Woodward, J. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

Yilmaz, L. 2012. "Reproducibility in m&s research: issues, strategies and implications for model development environments". *Journal of Experimental & Theoretical Artificial Intelligence* 24(4):457–474.

Yilmaz, L., and B. Liu. 2020. "Model credibility revisited: Concepts and considerations for appropriate trust". *Journal of Simulation*:1–14.

AUTHOR BIOGRAPHIES

LEVENT YILMAZ is the Alumni Distinguished Professor of Computer Science and Software Engineering at Auburn University. He holds M.S. and Ph.D. degrees in Computer Science from Virginia Tech, and B.S. degree in Computer Engineering from Bilkent University. His research interests are Theory and Methodology of Modeling & Simulation, Cognitive Computing, and Complex Adaptive Systems. He is a Fellow of the Society for Modeling and Simulation International (SCS) and is the founding organizer and General Chair of the Annual Agent-Directed Simulation Symposium series. His email address is yilmaz@auburn.edu.