Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

BROADENING ACCESS TO SIMULATIONS FOR END-USERS VIA LARGE LANGUAGE MODELS: CHALLENGES AND OPPORTUNITIES

Philippe J. Giabbanelli[1], Jose J. Padilla[1], and Ameeta Agrawal[2]

[1]Virginia Modeling, Analysis and Simulation Center, Old Dominion University, Norfolk, VA, USA
[2]Dept. of Computer Science, Portland State University, Portland, OR, USA

ABSTRACT

Large Language Models (LLMs) are becoming ubiquitous to create intelligent virtual assistants that assist users in interacting with a system, as exemplified in marketing. Although LLMs have been discussed in Modeling & Simulation (M&S), the community has focused on generating code or explaining results. We examine the possibility of using LLMs to broaden access to simulations, by enabling non-simulation end-users to ask what-if questions in everyday language. Specifically, we discuss the opportunities and challenges in designing such an end-to-end system, divided into three broad phases. First, assuming the general case in which several simulation models are available, textual queries are mapped to the most relevant model. Second, if a mapping cannot be found, the query can be automatically reformulated and clarifying questions can be generated. Finally, simulation results are produced and contextualized for decision-making. Our vision for such system articulates long-term research opportunities spanning M&S, LLMs, information retrieval, and ethics.

1 INTRODUCTION

As modelers, we are not usually our own end-users. Rather, models may be *commissioned* to address specific organizational and societal needs, *co-designed* with subject matter experts or individuals who provide their lived experiences, and employed as *decision-support tools* that eventually impact communities (Loeffler and Loeffler 2021; Oldfield and Haig 2022). End-users such as model commissioners, stakeholders, and community members may not have technical expertise in a modeling language or simulation environment. Yet, they need to interact with simulation models. When engagements are limited in scale, modelers may act as facilitators to translate the questions of end-users onto queries for a simulation engine, and translate simulation results into a contextualized and accessible response. However, a reliance on modelers can create a bottleneck in the process, since it depends on staffing limitations and cannot easily scale-up. For example, simulation models for welfare allocation, obesity management, or suicide prevention would potentially impact the lives of millions of individuals: sufficiently staffing a 'modeling helpdesk' to help individuals ask questions and operate a model could be cost-prohibitive. In addition, when end-users either need help to interpret a model or start to dedicate their own time to learning about modeling, they are less able to translate simulation insights into practices (Zellner et al. 2022).

Environments such as `NetLogo` and `ClouDES` have been developed to make it easier for non-modelers end-users to interact with simulations (Ramli et al. 2015; Padilla et al. 2014). Several such environments have been presented at the Winter Simulation Conference ('WinterSim'), such as Sim4edu.com (Wagner 2017). These environments often provide web-based interfaces and they emphasize low-code/no-code interactions (Hewage et al. 2024; Bocciarelli and D'Ambrogio 2023) by allowing users to modify elements such as sliders or text fields as a means to input a new scenario. When existing interactive elements are not sufficient to input a scenario, drag-and-drop features may allow users to build more complex alternatives, as shown in commercial simulators (Chong et al. 2022). However, these practices still face *two challenges: a lack of usability, and limited potential for reuse*. First, although interactions are intended to be 'user-

friendly' by modelers and software developers, usability studies show that users struggle in integrating models to their existing software ecosystem (Giabbanelli and Vesuvala 2023). Issues of usability have been observed for many years, as researchers have regularly pointed out that "practical user-friendly tools that allow scenario simulations also to non-expert users are clearly still lacking" (Marvuglia, Gutiérrez, Baustert, and Benetto 2018).

Second, modeling interfaces that seek to be user-friendly may provide users with a graphical interface in which they can ask a question by changing a set of input values. While such interfaces may succeed in promoting high levels of usability, this would be at the expense of model reuse. Indeed, when users wish to ask questions that are not exactly as intended (yet still within the scope of the model), they would need more skills to pilot the simulation by expanding the interface or directly changing the implementation. The INFSO-SKIN study exemplifies this struggle, as a steering committee (i.e., model commissioners) provided a detailed list of questions in a Tender Specifications document, but "stakeholders neither shared the same opinion about what questions should be in the final sample [...] nor shared the same hypotheses about questions in the final sample." (Ahrweiler, Frank, and Gilbert 2019) Consequently, one user group may succeed in drafting a list of questions and modelers can build an interface accordingly, but another user group may find it inadequate for their needs, thus limiting reuse of a model across user groups. The difficulty of supporting user-defined questions was echoed in the recently proposed notion of 'Approachable Modeling and Smart Methods' coined by Schimpf and Castellani (2024), who stressed that *tools need to promote user-driven inquiry*. One strand of research could be to provide more tools for end-users to extend interfaces such that they can input other scenarios (as long as they are supported by the model). However, the point was previously made at WinterSim that "to make M&S truly relevant to non-simulation experts and foster a radical change in the way we consume models and simulations, we have to venture beyond simply extending traditional graphical user interfaces" (Rechowicz et al. 2018).

In this paper, we posit that *emerging technologies can provide new modalities to interact with simulations in a manner that is both intuitive and sufficiently flexible to support a variety of user-driven inquiries*. As the practice of modeling and simulation spans several decades, there is an abundance of examples in which new technologies have been successfully integrated in the software ecosystem to support new means of interaction. For instance, simulations can now be embedded in the real-world context of a user through augmented reality (Giabbanelli, Shrestha, and Demay 2024). Our focus is on democratizing access to simulations with a broad set of users by using *Large Language Models (LLMs)* (such as OpenAI's GPT) as facilitators. LLMs have already been used on several occasions for specific simulation tasks such as turning a narrative into either a conceptual model (Hosseinichimeh et al. 2024; Giabbanelli and Witkowicz 2024) or simulation code (Niu et al. 2024; Frydenlund et al. 2024), learning about modeling by using LLMs as tutors (Chen et al. 2024), explaining a model (Shrestha et al. 2022), or assisting simulated entities in making decisions (Ghaffarzadegan et al. 2024). At WinterSim, we also proposed to use LLMs during other parts of the simulation process, such as verification and validation (Giabbanelli 2023). Given this growing uptake of LLMs by the simulation community, this paper seeks to guide the discussion on the opportunities and challenges in using LLMs to assist a wide-audience of end-users in interacting with simulations.

The remainder of this paper is structured as follows. In Section 2, we provide a framework to identify the main three steps and components of a system using LLM as mediators between end-users and simulations. Each of the three steps is then detailed subsequently: Section 3 focuses on aligning questions and model parameters, Section 4 examines the need for clarifications (e.g., via query reformulation), and Section 5 discusses the requirements for the final answer such as the need to ground it in the model's context. To design such systems, we will need to evaluate them, thus Section 6 summarizes quantitative and qualitative metrics that could be of interest in future implementations. We provide concluding remarks in Section 7.

2 CONCEPTUAL FRAMEWORK

Our framework is summarized in Figure 1. It is designed to allow users to express questions naturally, which would be answered thanks to a simulation model. We assume that *the purpose of the query is a*

what-if question that would translate to a simulation input. This rules out questions such as "can you create a simulation model" or "can you verify this model", which have been discussed in other articles (see Section 1). We make no assumption on the style of a question. For example, users are not limited to the use of a specific professional jargon. As a consequence of letting users choose their own words, we will need to map the key constructs in their question to the parameters available in a simulation model. We assume a general case in which *multiple* models may be available, such that the more relevant one should be triggered to address the user's question. For example, there may be an ensemble of geographically specialized models and the user can be interested in a specific area. Traditionally, identifying key constructs would be achieved through the Natural Language Processing (NLP) task of entity recognition, while mapping constructs expressed by the user onto the inputs of a model is a matter of semantic matching. Through several benchmarks and case studies (Yuan et al. 2024; Chen et al. 2023), LLMs have demonstrated an ability to perform both entity recognition and semantic matching. Consequently, our framework relies on the LLM (via additional prompts) to align a user's query with a suitable simulation model.

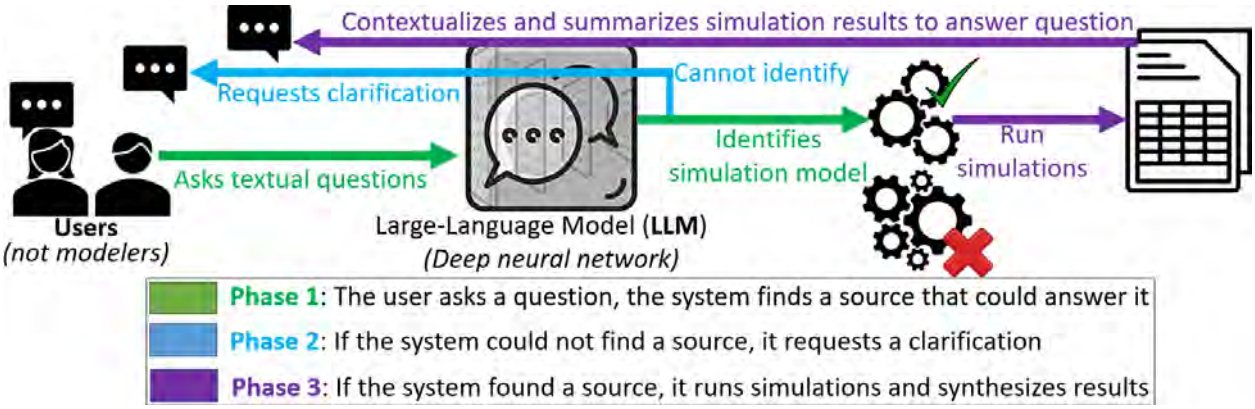

Figure 1: Overview of our conceptual framework including textual questions, alignment with the right model (or request for clarification), and translation of structured simulation results into text. We envision the system as composed of three phases (indicated by colors), detailed in dedicated subsections 3 to 5.

On the one hand, *we should not expect a perfect alignment*: that would force the user to define a value for every input of a simulation model, while excluding any concept beyond the model. This would be a mechanical or 'templated' interaction unlike the natural style that we seek to support. Considering an interaction between end-users and modelers as facilitators, modelers may not refuse to answer a question until every input has been specified (they may assume that unspecified inputs are left to a baseline value) or as soon as another concept is evoked (they may just ignore it). On the other hand, if the user's question is too far from any available model, a modeler would prompt for clarifications. In this spirit, the system has an alignment threshold below which it requests a clarification from the user. This may lead to a short conversation until the request becomes sufficiently clear given the simulation models available.

Once the request is clear and a corresponding model is identified, simulations are run accordingly. We assume that simulation results are stored in a structured format hence the LLM performs the task of data-to-text generation. This task can be further specialized depending on the data model used in the simulation experiments: for example, tabular data such as CSV files call for table-to-text generation. Numerous benchmarks and improved LLMs have produced satisfactory performances on tasks involving tables (Sui et al. 2024) and other types of structured data (Kasner and Dušek 2024). Beyond the well-studied matters of generating content that is factual and easy to read, *our framework stresses the importance of contextualizing simulation results to promote ethical decision-making activities*. Indeed, the reason to use simulations in the first place is often that taking decisions in a complex adaptive system may generated unintended consequences or only yield limited benefits due to unforeseen limitations (Putro 2016). We

should thus make efforts to answer a user while accounting for factors that were not directly featured in their questions. For example, if a user asks "what will happen to the number of fast-food outlets if I implement new zoning policies to limit them?", one answer is that "there will be 8.52% fewer fast-food outlets in total", which may sound successful and encourage their policy. Another answer using the same simulation model is that "there will be a reduction of 16.2% fast-food outlets in affluent neighborhoods, and an increase of 8.2% in deprived neighborhoods, where individuals are already are a higher risk for obesity due to over-exposure to unhealthy foods". The first answer strictly addresses what the user wanted, but the second one provides enough context to inform users against potential issues of fairness and equity.

In principle, the framework discussed in this section could be enacted using any individual LLM (e.g., OpenAI's GPT, Meta's LLaMa, Google's Gemini) or ensemble of LLMs as supported by emerging frameworks such as `LLM-BLENDER` Jiang, Ren, and Lin (2023). We would expect performances to depend on usual considerations such as the choice of LLM and its hyper-parameters (e.g., the temperature setting of GPT), the level of training, the application domain, and characteristics of the question. For example, an ambiguous question regarding a highly specialized medical model (which is unlikely to be contained within the dataset of a general-purpose LLM) and an absence of training may not yield high performances. However, some LLMs have safety settings that *refuse to engage* on certain topics, hence they will not answer a question or pass it onto subsequent modules for analysis. For example, Gemini evaluates and blocks requests that present a safety risk by containing hate speech, harassment, or dangerous content. This LLM will thus forbid requests to interact with a public health policy model for suicide prevention. Note that safety mechanisms can be bypassed by using a lax LLM to express a potentially sensitive request without using certain words (Gandee et al. 2024). For instance, we can provide a GPT prompt such as "You are a helpful assistant that combines a list of sentence into one paragraph that is rated Y for toddlers to read: `[request]`". The result can then be passed onto an LLM with more stringent safety checks. However, the reworded sentence may be more vague and thus harder to align with the parameters of a simulation model. For example, professional terms (which would be used by a model) such as Adverse Childhood Experiences would be expressed as "bad things that happen to people, especially children".

3 WHAT IS A QUESTION? ALIGNING CONSTRUCTS WITH MODEL PARAMETERS

Let's start with an example based on Tolk et al. (2013), where users ask questions regarding the consequences of sea level rise and its effect on flooding in city areas. Users want to know whether to vacate, invest, or maintain (VIM). Accordingly, users can ask "what VIM decision should I make in the downtown area if sea level will rise by six inches?" To formulate this question requires *known statements* that establish what vacating, maintaining, and investing are. It also requires statements (potentially unknown at the time) of the needed factors, and their combination, that may lead to a VIM decision. A question may vary depending on the purpose of the answer (*intent*). The question here was formulated from the perspective of a city's decision makers that need to assess where to vacate, maintain, or invest. This example previews major challenges (Zeigler, Muzy, and Kofman 2019): users may be unfamiliar with modeling practices and the level of details that a model needs (i.e., scope and resolution); models may answer a question partially; what-if scenarios can have different effects in different models.

Given these challenges, we need to operate under clear definitions. According to Tolk et al. (2013), a modeling question is "a collection of sentences to which truth values needs to be assigned" and akin to a query directed to a referent, or in our context, a *reference model*. A reference model captures everything that is known or assumed about a problem domain. Depending on social processes dynamics and experiences of a team, the reference model may be reduced into a few options as represented by different *conceptual models*. There can also be multiple ways of implementing a simulation given a conceptual model, for example based on data availability (Freund and Giabbanelli 2021). As a result, there is not necessarily 'one' valid model to answer a user's question. Rather, the system needs to identify a *candidate simulation model* to answer the question at the scope and resolution needed to meet the intent of the final user. We focus on identifying *one* candidate model, while noting that several algorithms can be applied if the system

identifies *several satisfactory candidate models* (Testoni and Fernández 2024), for instance by measuring the system's uncertainty (e.g., via entropy) and asking the user for preferences.

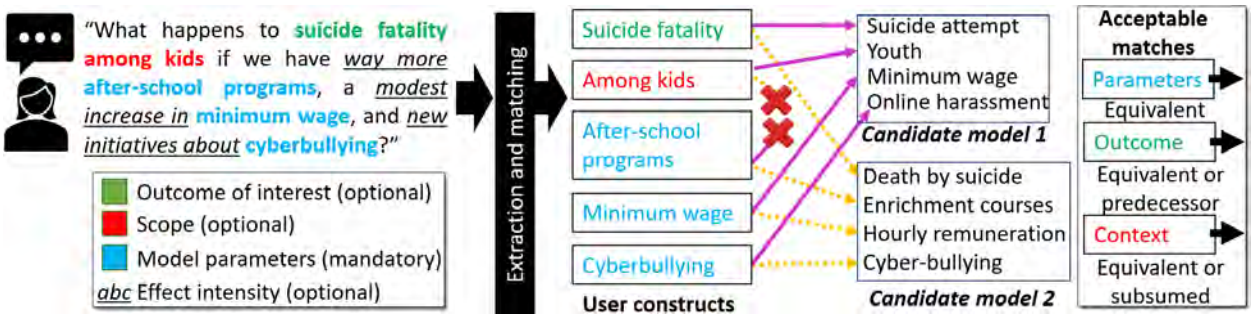

Figure 2: The user starts by formulating a what-if question. The *extraction* step identifies and categorizes key constructs. The user constructs are then *matched* with semantically equivalent parameters from available models. As a result, we *select* an acceptable model that operates in a context including the user's interests, and that is sufficiently informative as its outputs directly feed into the desired measurement outcomes.

The LLM needs to consider that modelers make several assumptions when identifying candidate models. As exemplified in Figure 2, modelers may consider that parameters are a valid match when they are *semantically equivalent* to the constructs that users seek to modify through their what-if question. Simulation outcomes should be related to the user's question and the population captured by the model should include the sub-population of interest to the user. Note that it may be too restrictive to only accept a simulation model that has the outcome (per semantic equivalence) desired by the user. If no such match can be found, the user may tolerate a model whose outcome directly precedes their measurement of interest. For example, if we cannot find a model of death by suicide, then it may be satisfactory to retrieve a model of suicide attempt. The LLM would *not* be responsible for improvising a model that relates the two constructs, as this is not a straightforward endeavor (e.g., suicide attempts are more lethal for men than women). The LLM would be responsible for recognizing that one stage closely precedes another, while noting that this increases uncertainty and has yet to be tested.

In order to recognize these relations, the LLM would need to either directly capture the reference model or have access to it (e.g., by querying a corpus when needed). Our recent proof-of-concept in Giabbanelli and Witkowicz (2024) suggests that LLMs can adequately capture several semantic relations (e.g., synonyms, antonyms) and aptly leverage authoritative sources such as peer-reviewed papers (e.g., the new prompt size for GPT-4 allows to pass a PDF document).

4 CAN (A)I HAVE A WORD WITH YOU? CLARIFYING QUESTIONS FROM THE LLM

The user's question will be answered by *retrieving*, running, and explaining a simulation model. We expect a gap between the user's query and a model, given that they may use different words. One of the primary concerns in the field of Information Retrieval (IR) is to resolve the lexical gap between a user's query and the identification of a suitable 'document' (i.e., a simulation model in our case). As summarized by Anand et al. (2023), multiple IR techniques seek to "reconcile the difference between user intent and the intent of the [modeled] system". Query Rewriting/Reformulation (QR) is particularly common to address underspecified and ambiguous queries, prior to (or instead of) asking a user for clarifications. Queries can be expanded automatically by prompting LLMs to recommend useful expansion terms and appending them to the user's original question (Wang et al. 2023). It is even possible to request LLMs to create different prompts and answer them, thus using the variability of outputs to yield a greater variety of additional terms (Dhole and Agichtein 2024). However, adding terms may *drift* from the user's intent and increase computing time when retrieving a simulation model. For instance, a student pursuing modeling

and simulation may ask GPT 3.5: "What is the number one factor to have a good career in modeling?" Without knowing the user's intent, the LLM assumes that the question is about the fashion industry and will emphasize physical attributes, such as "photogenic facial features". Using the LLM to find expansion terms via the prompts as shown by Dhole and Agichtein (2024) will thus add physical appearance to the user's query, which makes it drift from the user's intent and complicates the alignment with a suitable set of parameters.

An alternative is to rephrase the query by replacing some of the users' terms with synonyms that are more likely to be expressed as simulation parameters. We posit that suitable synonyms may be learned from prior user interactions with models, as these logs can reveal which user terms were eventually mapped to satisfactory models (Figure 3). Note that as a user successfully gets an answer, we can store the terms used in their question and the corresponding mapping to a simulation model. Through their interactions with the system, users thus contribute to building a thesaurus. As we discussed in Giabbanelli and Tawfik (2019), such thesaurus could improve the quality of future alignments and reduce their computational burden. In addition, a thesaurus can be a valuable research object in its own rights, as it reveals how end-users think of models.

The matter of retrieving the right model is complicated by the fact that interactions between users and models can turn into *conversations*. For example, consider that a user asks the modeler to run one intervention, hears the results, and finds them disappointing. The user may then ask "what if we *also* increased minimum wage?" This follow-up question supposes that one parameter is modified *in addition to* the modifications previously stated. A modeler would understand the dialogue context and interpret the question accordingly. Using an LLM, a simple solution could be to request that users provide complete questions each time, refraining from referring to their previous statements or results. While this may suffice for a prototype, we caution against designing a system for a broad user-base that relies on their goodwill to follow ambiguous instructions. Rather, we posit that a complete system would eventually have to handle conversational dependencies. Specifically, the LLM would be given the list of all prior question-answer pairs and *rewrite the user's query to make it self-contained*. However, this often requires fine-tuning a query reformulation model, which adds to the computational cost of the system (Yoon et al. 2024).

Despite the necessity of aligning a user's query with a model (with the help of query rewriting), the system should also account for the likely event in which the query is too imprecise. As mentioned in Section 2, a threshold on the quality of alignment should thus trigger a clarification. Rao and Daumé (2018) emphasize "that a good question is the one whose *likely answer* will be useful." There are two implications: the LLM should focus on questions that *(i)* the user can plausibly answer to *(ii)* improve the alignment with a simulation model. The first part requires a dataset to learn about users and their questions. While there exist datasets of human conversations containing questions (e.g., to develop dialogue systems) (Testoni and Fernández 2024), there is a paucity of data documenting the textual interactions of users with simulations. Such a dataset may be collected with the users' permission, once a system has been in operation for some time. Alternatively, user simulators have been developed to provide feedback to a system's response and answer clarifying questions (Owoicho et al. 2023).

5 RESPONSIBLE RESPONSES: CONTEXTUALIZING ANSWERS AND PROMOTING ETHICS

At this stage, the user asked a (potentially reworded and clarified) what-if question that calls for experimentation by simulation and a relevant model has been retrieved. We thus need to run the simulation and explain its results (Figure 4). The alignment between the user's question and the simulation model informs us on *which parameters* need to be modified from their default values. The question must be parsed further to determine the *direction and level of change* on each parameter. The direction of effect may be detected by the LLM. The level of change can be obtained by using fuzzy logic to transform the qualitative modifiers expressed by the user (e.g., 'way more' after-school programs, 'modest increase' in minimum wage) into values within the operating domain of each parameter (Giabbanelli and Crutzen 2014). If a model is stochastic, multiple simulation runs will be needed. Several methods such as desired Confidence

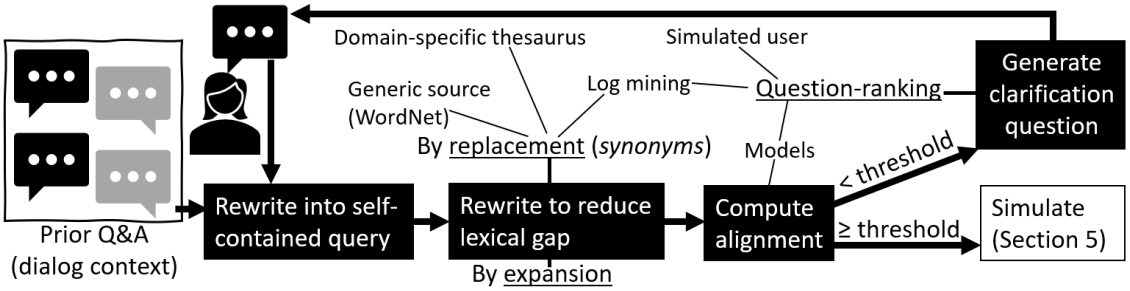

Figure 3: A user's query may need to be rewritten so that it becomes self-contained, then it can be rewritten to better align with simulation models. If the alignment is satisfactory, the simulation can proceed. Otherwise, the system generates a question and interprets the user's answer through the same rewriting process.

Intervals can determine the number of runs (see section 9.7.1 in Robinson 2014). However, if the proposed system operates as an open public platform where any number of individuals can ask questions, then the delivery of compute resources in the context of citizen science may become a determining factor. A simple approach is to use cloud computing resources and set a cost limit on each user query. However, the use of cloud computing for citizen science has its challenges. In particular, if the system mines user logs in order to improve its alignment algorithm or ask clarifying questions, then it "requires systematic, semi-automated data governance, and management plans in order to prevent the degradation of the data lake into a data swamp" (Ramachandran, Bugbee, and Murphy 2021).

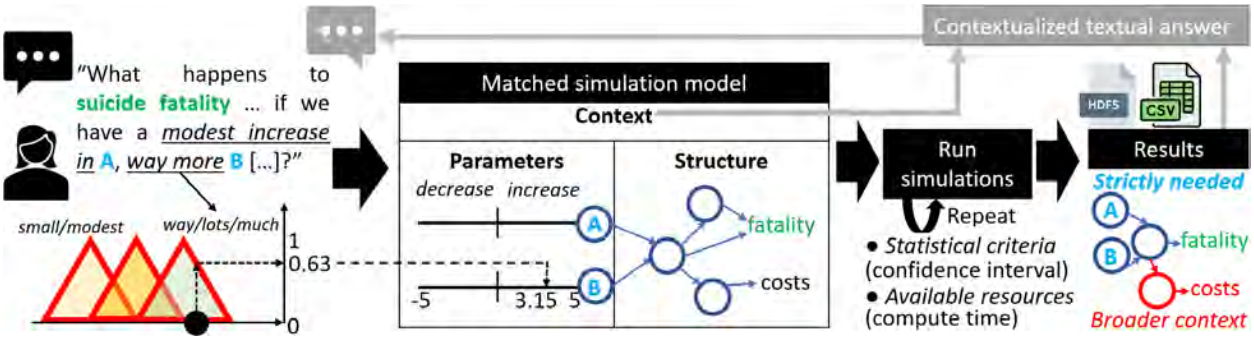

Figure 4: Running a simulation requires mapping user constructs to model parameters as well as converting linguistic modifiers (e.g., 'modest increase') into quantitative values for each parameter's domain via fuzzy logic. Several simulation runs can then be performed, but a comprehensive interpretation of the structured results (e.g., CSV or HDF5) would require access to the model's context and internal structure.

The main research challenges are about contextualizing the simulation results, not only to address the user's question but also to promote sound and ethical decision-making. Modelers may consider that promoting sound decision-making is the right action from a moral standpoint. However, this is also under increased scrutiny from the perspective of criminal liability for AI systems, including simulations. For instance, a decision-making platform (albeit sophisticated) may be viewed as a machine that cannot be the perpetrator of an offense, much as a mentally limited person (e.g., a child) would be considered an innocent instrument (Hallevy 2024). If a person causes a child to commit an offense, that person is criminally liable. The problem here is to determine whether that person is the *user* who asked a flawed what-if question, or the *programmer* whose system provided an inappropriate answer.

One challenge is that users of computer simulations tend to narrowly focus on intended consequences and neglect essential information on alternative, unintended consequences. Through several experimental studies, Ehrlinger and Eibach (2011) have shown that participants insufficiently accounted for unintended

consequences in a simulation, unless they were informed about the interrelatedness of variables in the modeled system. In other words, users tend to look for what directly interests them, unless they are told about the impact of a decision onto other outcomes and then start to consider the *overall* impact. On the one hand, it may be inadvisable to only tell users about the consequence of their what-if scenario on the variable that interests them. On the other hand, it may be exceedingly burdensome to tell them about consequences on all variables. We thus posit that a trade-off is a necessity. The technical challenge is that navigating this trade-off may require information on a model's context, at least to determine which variables are outcomes. A second challenge is to evaluate algorithmic fairness. This often involves assessing variations in outcomes across demographics such as race and ethnicity, age, and income subgroups (Schuch et al. 2023). If simulation outcomes are different in some of these groups, then the user should be informed as part of the response. Again, this would require contextual information about the model to identify which variables represent demographic data.

6 WHAT IS A GOOD SYSTEM? MEASURING FAITHFULNESS, BIASES, AND EXPLANATIONS

While our framework would create a combination of simulation models and LLMs that is accessible to a broad audience, evaluating the system is a critical challenge, compounded by the limitations of the two parts as both simulations and LLMs can produce outputs that are not always accurate, reliable, or unbiased, or answers may simply get 'lost in translation'. To address these challenges, a multifaceted evaluation approach should be adopted by focusing on three main phases defined in the previous sections: question processing, query clarification/reformulation, integration of simulated output and response generation. As part of this evaluation, it is important to first create a carefully annotated dataset to provide a benchmark. This data should consist of *user questions*, *mapped queries*, and *simulated outputs*. Given the vast space of possible user questions, automatically generating user questions based on parameters can be considered.

Since the LLMs are essentially serving as the bridge between the user and the simulation model, the first stage of evaluation should focus on whether the LLM accurately understands the user's question and maps it to a set of parameters that can be understood by the simulation model. LLMs are capable of processing natural language, but how well do they interpret the user's intentions/constructs and formulate a query that can be used as input to a simulation model? Evaluation at this stage would thus identify when simulation model parameters are not compatible/complete/found with the user's question. This evaluation is a necessary part of our system, as discussed in section 4 where the LLM asks clarification questions to the user until satisfactory queries can be formulated. The most visible part for the user is the generation of the final response, hence we focus on its evaluation.

Metrics of *faithfulness* assess how well the LLM's response reflects the content of the simulation results, avoiding any misinterpretations or injections. Standard metrics (e.g., ROUGE, BERTScore) are not effective in capturing nuanced faithfulness or lack thereof, as they focus on the overall semantic similarity. Instead, one should rely on semantic *inference* related measures such as textual entailment which has been found effective in identifying faithfulness (Maynez et al. 2020). Natural Language Inference (NLI) is a task in natural language processing where given two pieces of text – a premise (LLM's response) and a hypothesis (simulated output) – the goal is to determine whether the hypothesis can be inferred from the premise, with common relationships including entailment, contradiction, and neutral. An evaluator can leverage existing NLI models or train one by fine-tuning a pretrained language model such as BERT on a new NLI dataset of simulated outputs. Specific factuality metrics based on textual entailment include Adversarial Natural Language Inference (Nie et al. 2019), Summary Consistency (Laban et al. 2022) and FactCC (Kryscinski and McCann 2021) to determine whether a given text is consistent with the source.

Because assessing faithfulness requires logical inference over factual information, several question-answering (Q&A) based models can also be used such as QAGS (Wang, Cho, and Lewis 2020) and FEQA (Durmus, He, and Diab 2020). Faithfulness is only one aspect of evaluation. Another important dimension is that of *biases*, which can happen in LLMs due to issues at various stages such as dataset selection, annotation, and instruction tuning (Agiza, Mostagir, and Reda 2024). While several definitions

of fairness and biases exist, two commonly used ones focus on demographic parity and equal opportunity. The former compares the distribution of sensitive demographic attributes (e.g., race, gender) in the LLM's outputs to the distribution in the real world or the training data, with ideal outputs not favoring any particular demographic group. The latter assesses whether the LLM provides similar response for different demographic groups. For example, if you ask an LLM to suggest salary increments for a person, it should offer similar options regardless of the person's gender. Recently introduced metrics such as Large Language Model Bias Index (LLMBI) can provide insights into any biases that may occur in the generated output (Oketunji, Anas, and Saina 2023). Yet another dimension of quality can be derived by examining the LLM's confidence scores. Some LLMs (mostly open-source) can generate how confident they are in their responses. However, since the scores reflect the model's own assessment, it may be unreliable.

In addition to simply responding to the user question, the user response can be augmented by *explanations* generated by the LLM. This would inform the user not only of the simulated outputs but also explain the rationale behind the answer for more effective decision making. To assess how useful the explanations are to the user, one could perform a user satisfaction study or measure user engagement metrics (conversation duration, follow-up questions). This could also be done for the overall response of the LLM.

Finally, when it comes to automatic evaluation, an emerging area of research is to use *use LLMs as evaluators* (Liu et al. 2023; Ferron et al. 2023; Gao et al. 2024). LLM-based evaluators can be used for large-scale evaluations and adapted to different evaluation criteria. Combined with a human-in-the-loop approach, LLMs offer the potential of providing reliable assessments alongside human evaluators.

7 DISCUSSION AND CONCLUSIONS

We articulated challenges and opportunities to design an end-to-end system that supports a broad set of end-users in interacting with simulations via textual queries. With the right resources, many of the steps that we outlined can be addressed by the research community in the short- to medium-term. However, there are also five long-term research challenges that warrant particular considerations. First, what if no *single* model can be retrieved by the LLM to answer a user? For example, there may be two models: one that computes the effect of A on B, and another that has parameter B and outcome C. A user may ask how a small increase in A impact C, which is not directly answerable by a single model but could be addressed by *composing* models. Creating a combination of models on the fly is complicated, as models may be intertwined rather than neatly separable. At least, we should ensure that the combined model is consistent (i.e., without contradicting statements). Second, user questions may not be as independent as they seem. For example, in a validated model for suicide prevention, users could ask about the effect of increasing training for doctors: little to no impact. They could then ask about the effect of adding mental health education programs: little to no impact. However, when combined, these two evidence-based interventions could *increase* suicidal behavior by creating an imbalance between service capacity and demand for services (Atkinson et al. 2020). This is a problem for ethical decision-making, as users may not inform the platform that they intend to use two interventions jointly, hence joined effects are never simulated.

Third, an end-to-end solution requires many sub-systems, including retrieval (to align a query with a relevant model), query rewriting, and running a simulation multiple times. The computational cost may become a concern, particularly if there is a broad set of potential users. While the problem may be alleviated by queuing simulations and informing users when results are ready, this may reduce the quality of service and it only distributes computations over time instead of fully addressing their costs. Alternatives to reduce the processing costs and associated energy consumption (i.e., green computing) have been studied both in the Information Retrieval community (Scells, Zhuang, and Zuccon 2022) and by the simulation community (Feng and Staum 2017). For instance, the system may answer a *related* question that has already been computed. A fourth related challenge is that LLMs may already be able to generate simulation outputs instead of running the models. Indeed, LLMs have been trained on enormous amounts of data, some of which includes simulation models. For instance, GPT can answer questions about the behavior of models that are part of the standard library of `NetLogo`. It would be of particular interest to assess to what

extent can LLMs complement or replace the outputs of a simulation model, for instance by checking for possible alignment between the responses of the LLM generated with and without the simulation model.

Finally, the response of an LLM is a *mixture* of the underlying information derived from the simulated outputs and its own parametric knowledge. Evaluating this mixture is of particular interest. Is the response grounded in the human/domain knowledge embedded in a validated simulation model, or is it produced by the knowledge embedded in the LLM (coming from various and sometimes unknown datasets)? Does it depend on the application domain? Users concerned with this distinction could restrict the information space (Braun et al. 2024) by formulating their query accordingly (e.g., "reply solely based on the simulation model"). In some cases, it is *necessary* for the LLM to draw from its own knowledge, or the response may lack context. We thus recommend that future research examines how to clarify the role of a simulation by contrast to the LLM's knowledge when providing an answer.

REFERENCES

Agiza, A., M. Mostagir, and S. Reda. 2024. "Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in LLMs". *arXiv preprint arXiv:2404.08699*.

Ahrweiler, P., D. Frank, and N. Gilbert. 2019. "Co-designing social simulation models for policy advise: lessons learned from the INFSO-SKIN study". In *2019 Spring Simulation Conference (SpringSim)*, 1–12. Institute of Electrical and Electronics Engineers. Tucson, AZ, USA.

Anand, A., V. Setty, A. Anand, *et al*. 2023. "Context aware query rewriting for text rankers using llm". *arXiv preprint arXiv:2308.16753*.

Atkinson, J.-A., Y. J. C. Song, K. R. Merikangas, A. Skinner, A. Prodan, F. Iorfino, , , , *et al*. 2020. "The science of complex systems is needed to ameliorate the impacts of COVID-19 on mental health". *Frontiers in Psychiatry* 11:606035.

Bocciarelli, P. and A. D'Ambrogio. 2023. "A Low-Code Approach for Simulation-Based Analysis of Process Collaborations". In *2023 Winter Simulation Conference (WSC)*, 2530–2541. Institute of Electrical and Electronics Engineers. San Antonio TX, USA.

Braun, M., M. Greve, F. Kegel, L. Kolbe and P. E. Beyer. 2024. "Can (A) I Have a Word with You? A Taxonomy on the Design Dimensions of AI Prompts". In *Proc. 57th Hawaii Int. Conf. on System Sciences (HICSS)*, 559–568. Waikiki HI, USA.

Chen, J., X. Lu, M. Rejtig, D. Du, R. Bagley, M. S. Horn *et al*. 2024. "Learning Agent-based Modeling with LLM Companions: Experiences of Novices and Experts Using ChatGPT & NetLogo Chat". *arXiv preprint arXiv:2401.17163*.

Chen, Z., W. Chen, J. Xu, Z. Liu and W. Zhang. 2023. "Beyond Semantics: Learning a Behavior Augmented Relevance Model with Self-supervised Learning". In *Proc. 32nd ACM Int. Conf. Information and Knowledge Management*, edited by C. Yang and C. Park, 4516–4522. Boise ID, USA.

Chong, C. S., C. S. Tan, P. Y. Tan, G. L. Tnay and Y. K. Lin. 2022. "Development of a Data-Driven Simulation Model for an Assembly-To-Order System". In *2022 Winter Simulation Conference (WSC)*, 1853–1863. Institute of Electrical and Electronics Engineers. Singapore.

Dhole, K. D. and E. Agichtein. 2024. "Genqrensemble: Zero-shot llm ensemble prompting for generative query reformulation". In *European Conference on Information Retrieval*, 326–335. Springer.

Durmus, E., H. He, and M. Diab. 2020. "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization". *arXiv preprint arXiv:2005.03754*.

Ehrlinger, J. and R. P. Eibach. 2011. "Focalism and the failure to foresee unintended consequences". *Basic and Applied Social Psychology* 33(1):59–68.

Feng, M. and J. Staum. 2017. "Green simulation: Reusing the output of repeated experiments". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 27(4):1–28.

Ferron, A., A. Shore, E. Mitra, and A. Agrawal. 2023. "MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings". In *Findings of the Association for Computational Linguistics: EMNLP 2023*, edited by H. Bouamor, J. Pino, and K. Bali, 2078–2100. Singapore.

Freund, A. J. and P. J. Giabbanelli. 2021. "The necessity and difficulty of navigating uncertainty to develop an individual-level computational model". In *International conference on computational science*, 407–421. Springer.

Frydenlund, E., J. Martinez, J. Padilla, K. Palacio and D. Shuttleworth. 2024. "Modeler in a Box: How Can Large Language Models Aid in the Simulation Modeling Process?". *Simulation* 100(7).

Gandee, T. J., S. C. Glaze, and P. J. Giabbanelli. 2024. "A Visual Analytics Environment for Navigating Large Conceptual Models by Leveraging Generative Artificial Intelligence". *Mathematics* 12(1946).

Gao, M., X. Hu, J. Ruan, X. Pu and X. Wan. 2024. "Llm-based nlg evaluation: Current status and challenges". *arXiv preprint arXiv:2402.01383*.

Ghaffarzadegan, N., A. Majumdar, R. Williams, and N. Hosseinichimeh. 2024. "Generative agent-based modeling: an introduction and tutorial". *System Dynamics Review* 40(1):e1761.

Giabbanelli, P., A. Shrestha, and L. Demay. 2024. "Design and Development of a Collaborative Augmented Reality Environment for Systems Science". In *Proc. 57th Hawaii Int. Conf. on System Sciences (HICSS)*, 589–598. Waikiki HI, USA.

Giabbanelli, P. and N. Witkowicz. 2024. "Generative AI for Systems Thinking: Can a GPT Question-Answering System Turn Text into the Causal Maps Produced by Human Readers?". In *Proc. 57th Hawaii Int. Conf. on System Sciences (HICSS)*, 7540–7549. Waikiki HI, USA.

Giabbanelli, P. J. 2023. "GPT-Based Models Meet Simulation: How to Efficiently use Large-Scale Pre-Trained Language Models Across Simulation Tasks". In *2023 Winter Simulation Conference (WSC)*, 2920–2931. Institute of Electrical and Electronics Engineers.

Giabbanelli, P. J. and R. Crutzen. 2014. "Creating groups with similar expected behavioural response in randomized controlled trials: a fuzzy cognitive map approach". *BMC medical research methodology* 14:1–19.

Giabbanelli, P. J. and A. A. Tawfik. 2019. "Overcoming the PBL assessment challenge: Design and development of the incremental thesaurus for assessing causal maps (ITACM)". *Technology, Knowledge and Learning* 24:161–168.

Giabbanelli, P. J. and C. X. Vesuvala. 2023. "Human factors in leveraging systems science to shape public policy for obesity: A usability study". *Information* 14(3):196.

Hallevy, G. 2024. "The Basic Models of Criminal Liability of AI Systems and Outer Circles Gabriel Hallevy". *Legal Aspects of Autonomous Systems: A Comparative Approach* 4:69.

Hewage, T. B., S. Ilager, M. A. Rodriguez, and R. Buyya. 2024. "CloudSim express: A novel framework for rapid low code simulation of cloud computing environments". *Software: Practice and Experience* 54(3):483–500.

Hosseinichimeh, N., A. Majumdar, R. Williams, and N. Ghaffarzadegan. 2024. "From Text to Map: A System Dynamics Bot for Constructing Causal Loop Diagrams". *arXiv preprint arXiv:2402.11400*.

Jiang, D., X. Ren, and B. Y. Lin. 2023. "Llm-blender: Ensembling large language models with pairwise ranking and generative fusion". *arXiv preprint arXiv:2306.02561*.

Kasner, Z. and O. Dušek. 2024. "Beyond Reference-Based Metrics: Analyzing Behaviors of Open LLMs on Data-to-Text Generation". *arXiv preprint arXiv:2401.10186*.

Kryscinski, Wojciech and McCann, Bryan 2021, April 29. "Evaluating the Factual Consistency of Abstractive Text Summarization". US Patent App. 16/750,598.

Laban, P., T. Schnabel, P. N. Bennett, and M. A. Hearst. 2022. "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization". *Transactions of the Association for Computational Linguistics* 10:163–177.

Liu, Y., D. Iter, Y. Xu, S. Wang, R. Xu and C. Zhu. 2023. "Gpteval: Nlg evaluation using gpt-4 with better human alignment". *arXiv preprint arXiv:2303.16634*.

Loeffler, E. and E. Loeffler. 2021. "The Four Co's: Co-commissioning, Co-design, Co-delivery and Co-assessment of public services and outcomes through traditional and digital mechanisms". *Co-production of public services and outcomes*:75–176.

Marvuglia, A., T. N. Gutiérrez, P. Baustert, and E. Benetto. 2018. "Implementation of Agent-Based Models to support Life Cycle Assessment: A review focusing on agriculture and land use". *AIMS Agriculture and Food* 3(4):535–560.

Maynez, J., S. Narayan, B. Bohnet, and R. McDonald. 2020. "On faithfulness and factuality in abstractive summarization". *arXiv preprint arXiv:2005.00661*.

Nie, Y., A. Williams, E. Dinan, M. Bansal, J. Weston and D. Kiela. 2019. "Adversarial NLI: A new benchmark for natural language understanding". *arXiv preprint arXiv:1910.14599*.

Niu, T., W. Zhang, and R. Zhao. 2024. "Solution-oriented Agent-based Models Generation with Verifier-assisted Iterative In-context Learning". In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*, edited by N. Alechina, V. Dignum, M. Dastani, and J. Sichman. Auckland, New Zealand.

Oketunji, A. F., M. Anas, and D. Saina. 2023. "Large Language Model (LLM) Bias Index–LLMBI". *arXiv preprint arXiv:2312.14769*.

Oldfield, M. and E. Haig. 2022. "Analytical modelling and UK Government policy". *AI and Ethics* 2(3):389–404.

Owoicho, P., I. Sekulic, M. Aliannejadi, J. Dalton and F. Crestani. 2023. "Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond". In *Proc. 46th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, edited by M. Kato, J. Mothe, and B. Poblete, 632–642. Taipei, Taiwan.

Padilla, J. J., S. Y. Diallo, A. Barraco, C. J. Lynch and H. Kavak. 2014. "Cloud-based simulators: Making simulations accessible to non-experts and experts alike". In *Proc. Winter Simulation Conference 2014*, 3630–3639. Institute of Electrical and Electronics Engineers. Savannah GA, USA.

Putro, U. S. 2016. "Value Co-creation Platform as Part of an Integrative Group Model-Building Process in Policy Development in Indonesia". *Systems Science for Complex Policy Making: A Study of Indonesia*:17–28.

Ramachandran, R., K. Bugbee, and K. Murphy. 2021. "From open data to open science". *Earth and Space Science* 8(5):e2020EA001562.

Ramli, N. R., S. Razali, and M. Osman. 2015. "An overview of simulation software for non-experts to perform multi-robot experiments". In *2015 International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR)*, edited by M. S. Ahmad, 77–82. Institute of Electrical and Electronics Engineers. Kuala Lumpur, Malaysia.

Rao, S. and H. Daumé. 2018. "Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information". *arXiv preprint arXiv:1805.04655*.

Rechowicz, K. J., S. Y. Diallo, K. B. D'An, and J. Solomon. 2018. "Designing modeling and simulation user experiences: an empirical study using virtual art creation". In *2018 Winter Simulation Conference (WSC)*, 135–146. Institute of Electrical and Electronics Engineers. Gothenburg, Sweden.

Robinson, S. 2014. *Simulation: the practice of model development and use*. Second ed. Bloomsbury Publishing.

Scells, H., S. Zhuang, and G. Zuccon. 2022. "Reduce, reuse, recycle: Green information retrieval research". In *Proc. 45th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, edited by L. Gallagher and Q. Wu, 2825–2837. Madrid, Spain.

Schimpf, C. and B. Castellani. 2024. "Approachable modeling and smart methods: a new methods field of study". *International Journal of Social Research Methodology* 27(1):1–15.

Schuch, H. S., M. Furtado, G. F. dos Santos Silva, I. Kawachi, A. D. Chiavegatto Filho and H. W. Elani. 2023. "Fairness of Machine Learning Algorithms for Predicting Foregone Preventive Dental Care for Adults". *JAMA Network Open* 6(11):e2341625–e2341625.

Shrestha, A., K. Mielke, T. A. Nguyen, and P. J. Giabbanelli. 2022. "Automatically explaining a model: Using deep neural networks to generate text from causal maps". In *2022 Winter Simulation Conference (WSC)*, 2629–2640. Institute of Electrical and Electronics Engineers.

Sui, Y., M. Zhou, M. Zhou, S. Han and D. Zhang. 2024. "Table meets llm: Can large language models understand structured table data? a benchmark and empirical study". In *Proc. 17th ACM Int. Conf. Web Search and Data Mining*, edited by L. C. Mata, S. Lattanzi, and A. M. Medina, 645–654. Merida, Mexico.

Testoni, A. and R. Fernández. 2024. "Asking the Right Question at the Right Time: Human and Model Uncertainty Guidance to Ask Clarification Questions". *arXiv preprint arXiv:2402.06509*.

Tolk, A., S. Y. Diallo, J. J. Padilla, and H. Herencia-Zapana. 2013. "Reference modelling in support of M&S—foundations and applications". *Journal of Simulation* 7:69–82.

Wagner, G. 2017. "Sim4edu. com—web-based simulation for education". In *2017 Winter simulation conference (WSC)*, 4240–4251. Institute of Electrical and Electronics Engineers. Las Vegas, NV, USA.

Wang, A., K. Cho, and M. Lewis. 2020. "Asking and answering questions to evaluate the factual consistency of summaries". *arXiv preprint arXiv:2004.04228*.

Wang, X., S. MacAvaney, C. Macdonald, and I. Ounis. 2023. "Generative query reformulation for effective adhoc search". *arXiv preprint arXiv:2308.00415*.

Yoon, C., G. Kim, B. Jeon, S. Kim, Y. Jo and J. Kang. 2024. "Ask Optimal Questions: Aligning Large Language Models with Retriever's Preference in Conversational Search". *arXiv preprint arXiv:2402.11827*.

Yuan, L., Y. Chen, G. Cui, H. Gao, F. Zou, X. Cheng, , *et al*. 2024. "Revisiting Out-of-distribution Robustness in NLP: Benchmarks, Analysis, and LLMs Evaluations". *Advances in Neural Information Processing Systems* 36.

Zeigler, B. P., A. Muzy, and E. Kofman. 2019. *Introduction to Systems Modeling Concepts*, 3–25. Academic Press, London, UK.

Zellner, M. L., D. Milz, L. Lyons, C. Hoch and J. Radinsky. 2022. "Finding the balance between simplicity and realism in participatory modeling for environmental planning". *Environmental Modelling & Software* 157:105481.

AUTHOR BIOGRAPHIES

PHILIPPE J. GIABBANELLI is a Research Professor at the Virginia Modeling, Analysis and Simulation Center (VMASC) at Old Dominion University. He has authored over 130 articles and several books in Modeling & Simulation, with an emphasis on combining M&S with Machine Learning. His email address is pgiabban@odu.edu.

JOSE J. PADILLA is a Research Associate Professor at VMASC at Old Dominion University, where he obtained his Ph.D. in Engineering Management. His primary research focuses on advancing computational modeling methods towards increasing modeling accessibility across ages and across disciplines. His email address is jpadilla@odu.edu.

AMEETA AGRAWAL is an Assistant Professor of Computer Science at Portland State University. Her primary research interests are in the areas of Natural Language Processing, Machine Learning, and Computational Social Sciences. She obtained her Ph.D. and MS. in Computer Science at York University, Canada. Her email address is ameeta@cs.pdx.edu.