Anticipating the risks and benefits of counterfactual world simulation models

Lara Kirfel[*]
Department of Psychology
Stanford University
l.kirfel@stanford.edu

Robert J. MacCoun
Stanford Law School
Stanford University
rmaccoun@stanford.edu

Thomas Icard
Department of Philosophy
Stanford University
icard@stanford.edu

Tobias Gerstenberg
Department of Psychology
Stanford University
gerstenberg@stanford.edu

Abstract

This paper examines the transformative potential of *Counterfactual World Simulation Models* (CWSMs). CWSMs use pieces of multi-modal evidence, such as the CCTV footage or sound recordings of a road accident, to build a high-fidelity 3D reconstruction of the scene. They can also answer causal questions, such as whether the accident happened because the driver was speeding, by simulating what would have happened in relevant counterfactual situations. CWSMs will enhance our capacity to envision alternate realities and investigate the outcomes of counterfactual alterations to how events unfold. This also, however, raises questions about what alternative scenarios we should be considering and what to do with that knowledge. We present a normative and ethical framework that guides and constrains the simulation of counterfactuals. We address the challenge of ensuring fidelity in reconstructions while simultaneously preventing stereotype perpetuation during counterfactual simulations. We anticipate different modes of how users will interact with CWSMs and discuss how their outputs may be presented. Finally, we address the prospective applications of CWSMs in the legal domain, recognizing both their potential to revolutionize legal proceedings as well as the ethical concerns they engender. Anticipating a new type of AI, this paper seeks to illuminate a path forward for responsible and effective use of CWSMs.

1 Introduction

Imagine a pedestrian and a car collide on a busy intersection. Naturally, questions of responsibility and liability arise. Who is responsible for the collision, who is liable for the damage and injuries? Could the accident have been avoided, and if so, how? A CCTV camera recorded the last few seconds of the collision. However, this short clip alone won't contribute much to the clarification of the case. With the help of Artificial Intelligence (AI), this is about to change. Generative AI vastly expand the possibility of generating and interacting with evidence by building realistic reconstructions of what happened. Based on the CCTV footage, and a world model of car dynamics, street environments, and pedestrian behavior, AI-powered simulation models will soon be able to build a generative model of what happened and render a dynamic 3D simulation of how the crash came to pass [Gupta et al., 2020, Jadhav et al., 2020]. Such models will not only be able to reconstruct what happened, but also run

[*] *Corresponding author*: Lara Kirfel (l.kirfel@stanford.edu), Department of Psychology, 450 Jane Stanford Way, Building 420, Office 302, Stanford, CA 94305

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

counterfactual simulations of how things could have played out differently [see Tavares et al., 2021]. For example, the model's reconstruction from the CCTV footage might reveal that the driver was speeding. As users, we can then ask the question of whether the accident happened *because* the driver was speeding by simulating what would have happened if they hadn't. A model's counterfactual simulations of what would have happened if the driver hadn't been speeding lay the basis for a nuanced understanding of causality and promise to help evaluating questions of responsibility and liability.

An important goal in generative AI is to develop dynamic and accurate 3D environments, allowing for dynamic simulations consisting of objects, spaces, and agents [Kaur et al., 2023, Hu et al., 2023]. Here, we focus on a class of generative simulation models that we call "Counterfactual World Simulation Models" (CWSMs; see Figure 1). CWSMs represent an evolution from traditional image- and video-generating AI. CWSMs create digital replicas of real world scenarios based on different sources of evidence that can include images, video, audio, and text. CWSMs model the dynamic interaction of human agents in a physical environment over a limited period of time [Gan et al., 2020, Ivanovic et al., 2018, Brodeur et al., 2017, Clarke et al., 2022, Zhang et al., 2020, Li et al., 2018]. While world simulation models can be used to predict what will happen next [Cui et al., 2020, Sahoh et al., 2022], and to infer what happened in the past, *counterfactual* world simulation models can also simulate counterfactual scenarios of how things could have played out differently [Tavares et al., 2021, Feder et al., 2021, Gerstenberg and Stephan, 2021]. Currently, AI can synthesize realistic 3D environments [Kaur et al., 2023] and there are formal modeling approaches to counterfactual simulation [Tavares et al., 2021, Gerstenberg and Stephan, 2021, Gerstenberg, 2022, Bear et al., 2023], but there is not yet a single AI system that integrates evidence reconstruction and counterfactual simulation. However, the integration of these distinct capabilities into a comprehensive system is a natural progression in AI development. We therefore think the future use of CWSMs requires careful reflection.

Because counterfactual considerations about what would have happened are common practice in legal analysis and argumentation [Saxena et al., 2023], the application of generative AI could radically alter the landscape of legal proceedings [Alarie et al., 2018, Atkinson et al., 2020]. For example, we may ask whether an accident could have been avoided if the driver had driven more slowly. But how slowly exactly? What would have happened if the driver had behaved more reasonably, and how exactly would such "reasonable behavior" have looked like? Rather than referring to vague and speculative hypothetical scenarios, generative AI has the capability of providing vivid, detailed simulations to elaborate on these intricate questions. While the capabilities of CWSMs hold significant promise, their responsible deployment requires anticipating technological, social, and ethical challenges. Philosophers and psychologists have long grappled with questions surrounding what constitutes responsibility and liability, and how these concepts should be applied. With the advent of generative simulation models, these normative and descriptive questions will be thrust to the forefront of technological development. AI will pave the way from a single video frame of an accident to helping users find answers to questions like "What caused the accident?" and "Who is responsible?" via generative simulation.

In this paper, we first describe what CWSMs are and how users may interact with them. We discuss several ethical challenges that CWSMs face. Subsequently, we explore several prospective applications of these models within the legal sphere. This exploration includes an examination of their role in generating evidence by legal fact-finders, as well as the presentation and responsible use of such evidence in court.

2 Counterfactual world simulation models

Simulations can create matched digital worlds, allowing us to visualize and predict future scenarios or consider alternative outcomes. Large generative models, such as *Dall-E* or *StableDiffusion*, and recent advances in neural rendering, diffusion models, and attention architectures have paved the way for creating a new class of AI-powered simulators [Yuan and Veltkamp, 2021, Kapelyukh et al., 2023]. Text-to-3D and Image-to-3D generators convert a user's natural language commands or images into realistic 3D models. Instead of having to build models and animations by hand, generative AI digitally synthesizes 3D environments [Zhang et al., 2023, Gozalo-Brizuela and Garrido-Merchán, 2023] based on minimal input. And recent developments like the Metaverse even allow users to step inside these simulations by creating immersive virtual worlds[Surveswaran and Deshpande, 2023]

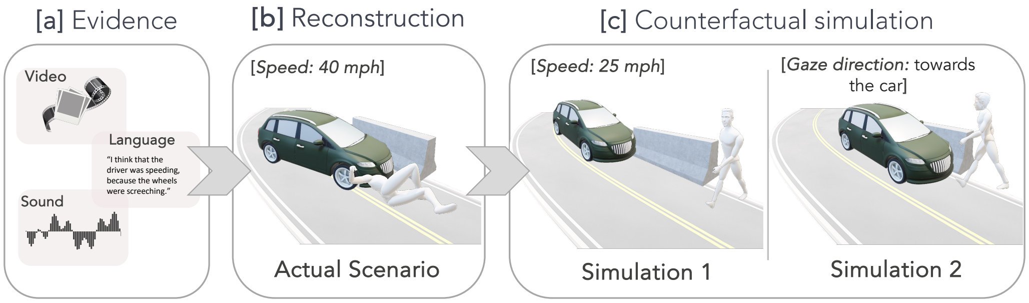
[a] Evidence
Video
Language
"I think that the driver was speeding, because the wheels were screeching."
Sound
[b] Reconstruction
[Speed: 40 mph]
Actual Scenario
[c] Counterfactual simulation
[Speed: 25 mph]
Simulation 1
[Gaze direction: towards the car]
Simulation 2

Figure 1: **Illustration of a Counterfactual World Simulation Model (CWSM) applied to a traffic accident.** The CWSM uses multi-modal evidence [a] to reconstruct what happened [b]. It can then simulate different counterfactual scenarios [c], to answer causal questions. For example, in Simulation 1, it considers what would have happened if the driver hadn't speeded. In Simulation 2, it considers what would have happened if the pedestrian had seen the car approaching. In both of these counterfactual simulations, the accident may have been avoided.

or simulate business decisions by creating virtual enterprises [Bian et al., 2021]. In the future, AI may create entire movies using text-to-video prompts [Ali et al., 2023]. The integration of causal and generative AI techniques allows users to pose counterfactual queries and expand the powers of simulation models even further: In addition to merely reconstructing evidence, generative AI can simulate what would have happened if certain factors in a situation would have been different [Tavares et al., 2021, Feder et al., 2021]. AI-assisted counterfactual simulation models provide powerful imagination machines that enable their users to entertain and depict "what-if" scenarios.

A CWSM operates in three stages. First, it processes multi-modal evidence which could encompass various forms of data including images, videos, sounds, and textual information ("Evidence", Figure 1a). Second, it reconstructs a detailed representation of what actually happened ("Reconstruction", Figure 1b; Baltrušaitis et al., 2018, Johnson et al., 2016). Third, it simulates what would have happened in relevant counterfactual scenarios ("Counterfactual Simulation", Figure 1c; Tavares et al., 2021). We describe each step in more detail below using the road accident example.

2.1 Evidence

What happened at the accident scene? In a first step to answer this question, an CWSM will need to integrate multi-modal pieces of evidence from the scene, weaving together strands of information from diverse sources ("Evidence", Figure 1a). Visual data like CCTV, dashcam or bike helmet footage, and potentially satellite and mobile phone imagery form the visual input to the model. Audio recordings or 911 calls inject sound dimensions to the model, offering insights into ambient and environmental conditions like traffic noise or mechanical failures. Textual data, verbal reports, comprising witness testimonies, police and medical reports etc., enrich the model with contextual conditions. The AI simulation model reconciles these disparate pieces of evidence and renders a coherent and detailed reconstruction of the accident [Singh et al., 2020].

2.2 Reconstruction

Based on limited CCTV footage and other pieces of evidence, CWSMs can then analyze the 2D images and videos to extrapolate depth information and reconstruct 3D scenes ("Reconstruction", Figure 1b). Embedded with a world model, CWSMs generate a simulation of what happened, and thereby infer, for example, the car's velocity at the scene. As a consequence, the CWSM might reveal that the driver was speeding. But how do we know that the model is right? If the CWSM will later help legal-fact finders draw inferences about liability and fault, users must trust it to display a high-fidelity to real-world systems from the very beginning [Jacovi et al., 2021, Yilmaz and Liu, 2022].

2.2.1 Validating the reconstruction accuracy

A CWSM that creates a realistic simulation from limited visual input will need to be validated against ground truth [Tolk et al., 2021]. For instance, if dashcam footage from several angles is available, the model might be trained to reconstruct the accident using all but one of the information sources and then validated against the remaining one to see if it can accurately reconstruct the held out data source. This process would then be rotated among all the footage sets in the case of *cross-validation* [Ghosh et al., 2020]. The AI model is trained on k-1 data subsets and validated on the remaining subset, with each subset used exactly once as the validation data.

Cross-verification from a variety of independent data sources can be used to validate the CWSM's output. Real-world data from traffic sensors, CCTV, GPS logs, body or bike helmet cams, vehicle detection sensors or traffic management systems can be used to compare the simulation output to what actually happened as captured by the different sources of evidence. For example, physical evidence from the accident scene, such as tire or skid marks, vehicle debris, and impact points, can be used to validate the AI's simulation. The simulation must also account for environmental factors like road wetness, traffic conditions and road layouts in a physically plausible way, for instance, by adjusting friction coefficients for wet roads.

An additional strategy to ensure a model's validity is to test its predictions of held-out future data based on past data. For example, the scene simulation model could be provided with the first t steps of the video evidence, and then predict how the remainder of the episode will unfold. This is where the model uses its understanding of physics, vehicle dynamics, and contextual information to extrapolate the future of the sequence. This prediction can then be compared against the video footage of what actually happened. Diverse validation metrics such as measures to capture image similarity (Structural Similarity Index) [Bakurov et al., 2022] or overlap between object bounding boxes (Intersection over Union) [Rezatofighi et al., 2019] can capture the prediction of specific variables, sets of variables, or the entire video sequence. Future reliance on scene simulation models in legal proceedings demands appropriate trust based on credible results [Yilmaz and Liu, 2022].

2.3 Counterfactual simulation

By synthesizing and reconstructing evidence, a CWSM can extrapolate information about aspects, objects or individuals in the scene that were likely present but not directly observed or recorded – for example, the driver's speed. Such a generative simulation model can then be used to simulate counterfactual scenarios [Tavares et al., 2021]. For example, given that the reconstruction showed that the driver was speeding, we might want to know if their speeding was actually what caused the accident to happen. CWSMs can accommodate a spectrum of causal inquiries that range from broad to specific. For example, a user could begin with a broad causal question ("What caused the outcome?"), exploring the key factors that contributed to an accident. Alternatively, a user could ask a specific causal question, such as whether the accident occurred because of the driver's speeding. The CWSM translates the causal query into a counterfactual prompt. When a user poses their causal question that corresponds to the counterfactual "Could the accident have been avoided if the car had not been speeding?", the CWSM needs to interpret and instantiate this counterfactual scenario [Lassiter, 2017b,a]. For example, it must decide how much slower the car should be going to evaluate the counterfactual. Should the car drive precisely at the speed limit, or even lower? Moreover, it needs to determine at what point in time the driver should have lowered its speed. Should the simulation test for the outcome in a scenario where the car reduces its speed right before the accident or at an earlier moment [Gerstenberg and Stephan, 2021, Von Kügelgen et al., 2023, Gerstenberg, 2022]?

In order to answer counterfactual queries, various specifics will have to be determined by the model. One way to resolve the ambiguity about how the counterfactual intervention should be realized is by changing the values of variables from their original value to the nearest possible value that renders the counterfactual intervention true [Karimi et al., 2020, Virgolin and Fracaros, 2023]. For example, the model would generate an alternative scenario in which the car's speed is reduced just until the speed criterion is met, but no further. Likewise, the model would simulate reduced speed only from when this limit comes into effect on the route, but not before. In the case of a lack of distinctive feature instantiations (e.g. when no clear speed limit is available), the CWSM can also generate a counterfactual scenario that would represent the most 'normal' scenario in this situation, with normality being a mix of the most statistically frequent behavior (e.g. average driving speed) and prescribed behavior (e.g. generally allowed or recommended driving speed; Bear and Knobe, 2017, Fazelpour, 2021). The CWSM might also consider multiple scenarios that fit the query, perhaps generating a distribution of outcomes based on different interpretations of "slower" or "driving below speed limit". For the term "slower", the model might consider various degrees of "slowness":—for example, 5, 10, and 15 mph slower than the actual speed at the time of the accident.

2.3.1 Constraints on counterfactual simulations

We might not only be interested in the causal role of the driver, but also whether the pedestrian had any fault in the incident. Could the pedestrian have behaved differently such that the accident would have been avoided? On an ethical level, the general question arises what simulations should be admissible [Fazelpour, 2021]. AI-assisted simulation models will be able to generate any kind of variation of an actual scene and produce all kinds of behavior, circumstances, and sequence of events. However, from an ethical perspective, not everything that is possible is also admissible. This is especially true when it comes to the simulation of human behavior. What kind of alternative actions are appropriate to compare an agent's actual behavior against?

2.3.1.1 De-biasing simulations We suggest that there should be constraints on the simulation of counterfactual scenarios that acknowledge ethical dimensions of simulating agent behavior. Unlike when simulating alternative physical events, we think that simulations of how an agent could have behaved differently must meet standards of personal and cultural appropriateness, and not be unreasonable [Gardner, 2015]. For example, mobility limitations, visual, hearing or cognitive impairments must be taken into consideration [Leo and Goodwin, 2016, Francillette et al., 2020]. At the same time, simulations should avoid perpetuating stereotypes or biased narratives associated with certain groups (e.g., always portraying a specific racial group as jaywalkers or aggressive drivers, a certain gender as driving more aggressively, or portraying cheaper cars as being driven carelessly). Visual generative AI has been shown to significantly underperform when tasked with generating detailed images about minority groups [Mbalaka, 2023]. Likewise, image generation models are prone to simulate visualizations based on racial or socioeconomic stereotypes that exist in publicly available images. If a text prompt references a certain concept, such as a social group, the model must infer the main characteristics and fill in all the information that is underspecified. Text-to-Image AI might hence render neutral genderless language into visually biased representations [Bianchi et al., 2023, Luccioni et al., 2023]. Simulation models often employ humanoid avatars. The generation of 3D scenes and sequential human behavior – adding time, motion, audio – offers seemingly endless degrees of freedom to fill in underdetermined gaps. Hence, the simulation might not only adopt visual stereotypes; it might also incorporate stereotypical behaviors or actions associated with the biased representation. The richness and dynamics of 3D simulations can in fact amplify the problem of stereotyping in AI-generated simulations.

To ensure responsible and ethical use of CWSMs, we suggest implementing user-side guardrails that restrict what types of simulations a user can run. Algorithmic fairness audits can help identify and rectify biases [De Schutter and De Cremer, 2023]): Are the alternative behaviors being simulated equitable across different groups? Are counterfactual scenarios for pedestrian people of color or women more likely to be causing the accident by e.g. reckless behavior? While high representational fidelity of an agent might be desirable at the stage of reconstructing the evidence, it might lead to bias at the stage of simulating counterfactuals. In the initial phase of reconstructing the evidence, it's crucial for the AI to have high representational fidelity. For example, it should generate a simulation that accurately depicts a teenager walking across the street while looking at a smartphone. However, in order to simulate a scenario to answer the question, "Would the accident have been avoided if the pedestrian had behaved differently?", because of the high representational fidelity from the reconstruction phase, the CWSM might automatically associate distracted behavior and rely on the stereotype that all young individuals are constantly absorbed in their devices and careless about their surroundings. This could lead the CWSM to generate a counterfactual scenario that overly emphasizes their carelessness in other aspects, thus perpetuating the stereotype that younger individuals are generally inattentive and reckless. We therefore recommend the use of abstract or simplified visual representations of agents, involving generic, featureless avatars at the counterfactual simulation stage. For example, clothing, hairstyles, and accessories can be kept simple and generic, avoiding any culturally specific, age-related, or gendered cues. In cases where macro behaviors like pedestrian paths are in focus, simulations can represent agents as abstract symbols or icons that convey their movements and interactions.

2.3.1.2 Restricting counterfactual simulation of agent behavior The question of how a driver or pedestrian should have behaved differently should also be restricted by cultural and contextual considerations: We recommend that simulations should be sensitive to – and not go beyond – the norms, regulations, and practices specific to the location where the scenario is being simulated. Could the pedestrian have behaved differently, and if so how? An AI-generated simulation will need to be proportionate with regards to varying parameters like walking speed, reaction time and awareness of surroundings, and contextually sensitive to crossing behavior, pedestrian right-of-way, or in case of car drivers, lane discipline, driving etiquette or horn usage, etc. [Solmazer et al., 2020]. For example, expecting a pedestrian to predict the precise movements of a speeding vehicle may be beyond human capability and should not be considered a reasonable counterfactual. Likewise, the counterfactual behaviors a CWSM simulates should align with what is known about human decision-making under duress or in quick-response situations. Extreme and implausible behaviors, such as suddenly swerving across the double yellow line, even when it would potentially avoid running over a pedestrian [Goode, 2009] is potentially prejudicial and should require explicit and careful justification [Jager and Janssen, 2002, Suo et al., 2021].

3 Interacting with a counterfactual world simulation model

Design and accessibility of a CWSM system plays a pivotal role in shaping user interactions and the kind of counterfactual scenarios they will probe. As inputs are manipulated, CWSMs will offer visual feedback on the potential outcomes of the counterfactual scenarios. Here we briefly discuss the ways in which users may interact with CWSMs, focusing on the inputs they would provide to the model, and the outputs they would receive.

3.1 Model input

The way a counterfactual prompt is given can significantly affect the counterfactual simulation. Natural language prompts that are typed into a text box are more intuitive and make the system accessible to non-experts. However, they are also prone to ambiguities since natural language is often less precise than formal or coded language [see, e.g., Lassiter, 2017a,b]. While coded prompts or parameter-setting interfaces offer more precise control over the counterfactual conditions they require a deeper understanding of how the model works [Bove et al., 2023]. Such interfaces would allow fine-grained control over simulations through sliders, dropdowns, and numerical inputs; for example, setting car speed to "35 mph", reaction time to "1.5 sec", etc. Adequate Graphical User Interfaces can help users to pose complex queries using a drag-and-drop interface where they can select variables, conditions, and outcomes from dropdown menus and link them together to form a query [Jusiega, 2022].

3.2 Model output

After having established which counterfactual simulations a CWSM should generate, there is a question as to how its output should be presented. Given that there is more than one possible simulation of how a scenario could have occurred, how should the model communicate its uncertainty? Let's say that a CWSM has produced several different simulations for one scene. Should all of them be reported? Only the best? And what if they are all relatively "good" in terms of explaining the outcome, but are very different from each other [Yacoby et al., 2022]? The fact that these systems may produce multiple simulations with varying degrees of probability is of course advantageous for capturing the nuances of a situation, but it also complicates their interpretability [Goode, 2009].

In general, recipients and end-users of AI-assisted counterfactual simulations of any kind may not be aware of the error or uncertainty involved in reconstructing the scene, and hence may subconsciously be biased toward a strong belief in a simulated reconstruction [Ma et al., 2010]. How can we augment their presentation to facilitate a nuanced understanding? We suggest the sequential presentation of a subset of simulations, in order of their likelihood, each labeled with uncertainty indices. But which subset? Ideally, the selection of simulations should meet the following criteria: When they display scenarios in which the outcome changes, they are sparse, that is, they contain a minimal number of changes needed to change the outcome. Research in explainable AI suggests that the most effective counterfactual explanations often involve the smallest change necessary to alter the decision [Guidotti, 2022]. In addition, counterfactual simulations should be plausible and feasible, modifying features that make sense to users (e.g., driver's behavior vs. changes in traffic laws or structure of the street environment) and adhere to ethical guidelines (see above). They should also be diverse, including a variety of feature changes to offer alternative scenarios that highlight different perspectives [Smyth and Keane, 2022]. The simulation results should be effectively displayed in a visual interface where feature changes in the simulations are highlighted and interactive methods are provided for users to explore the data and model [Gathani et al., 2021, Shneiderman, 2020].

4 Applications of counterfactual world simulation models

An AI system capable of integrating pieces of evidence into 3D world models and simulating the outcome of an alternative course of events can be applied in a variety of domains. For example, in sports analysis, counterfactual simulations based on video evidence can be used to analyze whether different team tactics or individual players' actions would have led to a different outcome [Rahimi et al., 2019, Nakahara et al., 2022, Fujii et al., 2022]. Mobile home assistants equipped with a CWSM will be able to infer what happened and why based on different pieces of evidence, and take appropriate actions such as investigating further and informing the home owner [Suryavamsi and Arockia Selvakumar, 2019, Lopez et al., 2017]. Here, we focus on how CWSMs applications in the law. AI-powered counterfactual simulations can serve as a robust foundation for legal arguments, potentially transforming how evidence is generated and presented in legal contexts [Pereira et al., 2023]. However, their application also raises ethical and procedural concerns.

4.1 CWSMs as tools for aiding legal fact finders

Simulations can provide a powerful and persuasive tool in legal court cases. In consequence, they need to meet several restrictions and regulations to ensure fair and ethical use [Schofield, 2009]. When it comes to the question of how an agent could have behaved differently in a certain scenario, not all simulations that are physically possible and ethically sound should be allowed as evidence in court. What kind of simulations should legal fact-finders be allowed to share in a hearing, and which ones are inappropriate for consideration in legal proceedings?

Suppose the car accident from the beginning is turned into a personal injury case. The defendant's legal team employs AI-powered generative simulations that display a variety of possible alternative behaviors of the plaintiff in which no accident occurs, each of them highlighting the contribution of the plaintiff's own actions to the accident. Simulations can generate counterfactuals that do make a difference to the outcome (e.g., where the accident does not occur), as well as ones that do not make a difference (e.g., where the accident still occurs). Legal fact-finders can make the case for how a certain change in an agent's behavior would or would not have avoided the outcome in question.

Simulations should not unjustly portray the agent as at fault or as acting negligently or recklessly without appropriate supporting evidence [Shults et al., 2018]. Merely simulating behavior without appropriate supporting evidence can create an unjust and biased perception of guilt, bypassing the need for a thorough examination of evidence and due process. We recommend that attorneys should be allowed to generate both "upward" and "downward counterfactuals" [Roese, 1994]. An upward counterfactual is a scenario in which a change in the past leads to a more desirable or better outcome in the present, while in a downward counterfactual a change in the past leads to a worse or less desirable outcome in the present. However, special restrictions apply to simulations that display a defendant's behavior as significantly worse than it is. While downward counterfactuals can be applied as a strategic move to position the defendant's actual behavior in a more favorable light, they must meet the above outlined standards of reasonableness and fairness. Likewise, repeated exposure to simulations of harmful outcomes can diminish emotional impact that such events would normally elicit, and be used as a tool for desensitizing jury members towards a harmful actions [Williams and Jones, 2005].

The integration of AI-generated counterfactual simulations into legal proceedings will also present a transformative shift in how the "reasonable person" standard is interpreted and applied. The reasonable person standard is a legal construct used to evaluate the legality of an individual's actions by comparing them to what a hypothetical "reasonable person" would have done under similar circumstances [Tobia, 2018, Gerstenberg et al., 2018, Green, 1967, Wu and Gerstenberg, 2023, Lagnado and Gerstenberg, 2017]. Traditional legal frameworks rely on abstract, often generalized, notions of what constitutes "reasonable" behavior in various situations [Gardner, 2015]. However, the application of AI simulations allows for a far more detailed, granular and vivid depiction of alternative behaviors. How would a reasonable pedestrian have behaved, what speed would they have walked, what path, where would they have stopped, where would they have looked? The rich detail of generative AI forces legal professionals to redefine what behaviors count as "reasonable" in a much more precise and contextual manner. By running a variety of simulations, AI can generate a distribution of "reasonable" behaviors to specific situations. Such a distribution would show the range and likelihood of various actions that could be considered reasonable under the given circumstances, and allow to compare and quantify the "reasonableness" of the actual behavior.

4.2 CWSMs as tools for presenting evidence in court

At the end of a counterfactual simulation cycle stands the potential to use of its results as expert testimony in legal proceedings. The persuasive power of realistic simulations as evidence in court, however, poses challenges for procedural fairness. Viewers may an uncritically believe the presented material [Clifford and Kinloch, 2008]. Simulation models can provide visual representations of complex events or processes, making them more accessible and understandable to judges, jurors, and other stakeholders [Lagnado, 2021]. However, the realistic rendering of components of the virtual model may issue in a kind of "seeing-is-believing" effect [Ma et al., 2010, Etienne, 2021].

We advocate for clear guidelines of using AI simulations as evidence by proposing limitations on the level of detail for simulation objects, agents, and environments. Simulations should display high physical fidelity but otherwise display a low level of detail when it comes to the simulation of agents. For example, visual elements used in the simulation can be designed to be neutral and devoid of any gender-specific or otherwise individualized attributes. Facial expressions, being powerful conveyors of emotion and intent, should be handled with extreme care in simulations to avoid unintentional misrepresentations [Niedenthal et al., 2010]. This will minimize over-persuasive powers [Ma et al., 2010] and respect agents' dignity. We also suggest a framework for *counter-modeling* and resolving conflicting interpretations. Counter-modeling refers to the practice of developing alternative simulation models or interpretations that challenge the findings or conclusions of a particular simulation model. In legal proceedings, opposing parties may employ counter-modeling to present conflicting simulation results, aiming to support their own arguments or cast doubt on the reliability of the other team's simulation.

Consider again the aforementioned personal injury case. The plaintiff's expert witness presents a simulation model that suggests the defendant's vehicle was speeding at the time of the accident, leading to severe injuries for the plaintiff. However, the defendant's legal team employs counter-modeling by developing an alternative simulation model that incorporates different assumptions and factors. Their model suggests that the plaintiff's own actions contributed to the accident. We suggest guidelines that allows prosecution and defense equal access to the model, both in private preparation as well as in court. We propose that parties run their opponents' simulations with alternative assumptions that are of interest for their own argumentation, showing fact-finders how the model behaves under certain counterfactual conditions [Ma et al., 2010]. By employing counter-modeling, parties in legal proceedings can engage in a more robust and comprehensive evaluation of simulation models. This approach promotes a deeper understanding of the underlying assumptions, enhances transparency, and facilitates the resolution of conflicting interpretations.

5 Conclusion: A new kind of AI

The ability for counterfactual reasoning, that is, reasoning about alternative scenarios and speculating on what could have happened under different conditions, is fundamental for human reasoning, decision-making, and intelligence. AI-generated counterfactual simulations will radically expand our ability to imagine alternative realities and explore the consequences of hypothetical changes to the course of events. Generative AI as a tool to play with reality, however, comes with the responsibility for considerate application and the ethical use of the generated insights and knowledge. We have outlined some of the normative, practical and legal challenges, and offered some proposals for how to respond to them. In particular, we have focused on how the generation of counterfactual simulations will change the way we gather evidence and adjudicate in legal proceedings. Our paper provides an overview and guide to some of the most pressing and distinctive issues for this new kind of AI.

References

B. Alarie, A. Niblett, and A. H. Yoon. How artificial intelligence will affect the practice of law. *University of Toronto Law Journal*, 68(supplement 1):106–124, 2018.

S. Ali, D. DiPaola, R. Williams, P. Ravi, and C. Breazeal. Constructing dreams using generative ai. *arXiv preprint arXiv:2305.12013*, 2023.

K. Atkinson, T. Bench-Capon, and D. Bollegala. Explanation in ai and law: Past, present and future. *Artificial Intelligence*, 289:103387, 2020.

I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi. Structural similarity index (ssim) revisited: A data-driven approach. *Expert Systems with Applications*, 189:116087, 2022.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

A. Bear and J. Knobe. Normality: Part descriptive, part prescriptive. *cognition*, 167:25–37, 2017.

D. M. Bear, K. Feigelis, H. Chen, W. Lee, R. Venkatesh, K. Kotar, A. Durango, and D. L. Yamins. Unifying (machine) vision via counterfactual world modeling. *arXiv preprint arXiv:2306.01828*, 2023.

Y. Bian, J. Leng, and J. L. Zhao. Demystifying metaverse as a new paradigm of enterprise digitization. In *International Conference on Big Data*, pages 109–119. Springer, 2021.

F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023.

C. Bove, M.-J. Lesot, C. A. Tijus, and M. Detyniecki. Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 188–203, 2023.

S. Brodeur, E. Perez, A. Anand, F. Golemo, L. Celotti, F. Strub, J. Rouat, H. Larochelle, and A. Courville. Home: A household multimodal environment. *arXiv preprint arXiv:1711.11017*, 2017.

S. Clarke, N. Heravi, M. Rau, R. Gao, J. Wu, D. James, and J. Bohg. Diffimpact: Differentiable rendering and identification of impact sounds. In *Conference on Robot Learning*, pages 662–673. PMLR, 2022.

M. Clifford and K. Kinloch. The use of computer simulation evidence in court. *Computer Law & Security Review*, 24(2):169–175, 2008.

P. Cui, Z. Shen, S. Li, L. Yao, Y. Li, Z. Chu, and J. Gao. Causal inference meets machine learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3527–3528, 2020.

L. De Schutter and D. De Cremer. How counterfactual fairness modelling in algorithms can promote ethical decision-making. *International Journal of Human–Computer Interaction*, pages 1–12, 2023.

H. Etienne. The future of online trust (and why deepfake is advancing it). *AI and Ethics*, 1(4): 553–562, 2021.

S. Fazelpour. Norms in counterfactual selection. *Philosophy and Phenomenological Research*, 103 (1):114–139, 2021.

A. Feder, N. Oved, U. Shalit, and R. Reichart. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021.

Y. Francillette, E. Boucher, N. Bier, M. Lussier, K. Bouchard, P. Belchior, and S. Gaboury. Modeling the behavior of persons with mild cognitive impairment or alzheimer's for intelligent environment simulation. *User Modeling and User-Adapted Interaction*, 30:895–947, 2020.

K. Fujii, K. Takeuchi, A. Kuribayashi, N. Takeishi, Y. Kawahara, and K. Takeda. Estimating counterfactual treatment outcomes over time in complex multi-agent scenarios. *arXiv preprint arXiv:2206.01900*, 2022.

C. Gan, J. Schwartz, S. Alter, D. Mrowca, M. Schrimpf, J. Traer, J. De Freitas, J. Kubilius, A. Bhandwaldar, N. Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.

J. Gardner. The many faces of the reasonable person. *Law Quarterly Review*, 131(1):563–584, 2015.

S. Gathani, M. Hulsebos, J. Gale, P. J. Haas, and Ç. Demiralp. Augmenting decision making via interactive what-if analysis. *arXiv preprint arXiv:2109.06160*, 2021.

T. Gerstenberg. What would have happened? counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society B*, 377(1866):20210339, 2022.

T. Gerstenberg and S. Stephan. A counterfactual simulation model of causation by omission. *Cognition*, 216:104842, 2021.

T. Gerstenberg, T. D. Ullman, J. Nagel, M. Kleiman-Weiner, D. A. Lagnado, and J. B. Tenenbaum. Lucky or clever? From expectations to responsibility judgments. *Cognition*, 177:122–141, 2018. ISSN 00100277. doi: 10.1016/j.cognition.2018.03.019.

S. Ghosh, W. Stephenson, T. D. Nguyen, S. Deshpande, and T. Broderick. Approximate cross-validation for structured models. *Advances in Neural Information Processing Systems*, 33:8741–8752, 2020.

S. Goode. The admissibility of electronic evidence. *Rev. Litig.*, 29:1, 2009.

R. Gozalo-Brizuela and E. C. Garrido-Merchán. A survey of generative ai applications. *arXiv preprint arXiv:2306.02781*, 2023.

E. Green. The reasonable man: Legal fiction or psychosocial reality? *Law & Society Review*, 2: 241–258, 1967.

R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

S. Gupta, M. V. Sharma, and P. Johri. Artificial intelligence in forensic science. *International Research Journal of Engineering and Technology*, 7(5):7181–7184, 2020.

A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

B. Ivanovic, E. Schmerling, K. Leung, and M. Pavone. Generative modeling of multimodal multi-human behavior. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3088–3095. IEEE, 2018.

A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

E. B. Jadhav, M. S. Sankhla, and R. Kumar. Artificial intelligence: Advancing automation in forensic science & criminal investigation. *Journal of Seybold Report ISSN NO*, 1533:9211, 2020.

W. Jager and M. Janssen. The need for and development of behaviourally realistic agents. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, pages 36–49. Springer, 2002.

J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

V. Jusiega. *Designing a User Interface for Counterfactual Simulations of Adaptive Treatment Strategies*. PhD thesis, Massachusetts Institute of Technology, 2022.

I. Kapelyukh, V. Vosylius, and E. Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics. *IEEE Robotics and Automation Letters*, 2023.

A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera. Towards causal algorithmic recourse. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 139–166. Springer, 2020.

D. P. Kaur, N. P. Singh, and B. Banerjee. A review of platforms for simulating embodied agents in 3d virtual environments. *Artificial Intelligence Review*, 56(4):3711–3753, 2023.

D. A. Lagnado. *Explaining the Evidence: How the Mind Investigates the World*. Cambridge University Press, 2021.

D. A. Lagnado and T. Gerstenberg. Causation in legal and moral reasoning. In M. Waldmann, editor, *Oxford Handbook of Causal Reasoning*, pages 565–602. Oxford University Press, 2017.

D. Lassiter. Complex antecedents and probabilities in causal counterfactuals. In *Proceedings of the 21st Amsterdam Colloquium*, pages 45–54, 2017a.

D. Lassiter. Probabilistic language in indicative and counterfactual conditionals. In *Semantics and linguistic theory*, volume 27, pages 525–546, 2017b.

J. Leo and D. Goodwin. Simulating others' realities: Insiders reflect on disability simulations. *Adapted physical activity quarterly*, 33(2):156–175, 2016.

Y. Li, Z. Cui, Y. Liu, J. Zhu, D. Zhao, and J. Yuan. Road scene simulation based on vehicle sensors: An intelligent framework using random walk detection and scene stage reconstruction. *Sensors*, 18 (11):3782, 2018.

A. Lopez, R. Paredes, D. Quiroz, G. Trovato, and F. Cuellar. Robotman: A security robot for human-robot interaction. In *2017 18th International Conference on Advanced Robotics (ICAR)*, pages 7–12. IEEE, 2017.

A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

M. Ma, H. Zheng, and H. Lallie. Virtual reality and 3d animation in forensic visualization. *Journal of forensic sciences*, 55(5):1227–1231, 2010.

B. Mbalaka. Epistemically violent biases in artificial intelligence design: the case of dalle-e 2 and starry ai. *Digital Transformation and Society*, (ahead-of-print), 2023.

H. Nakahara, K. Takeda, and K. Fujii. Estimating the effect of hitting strategies in baseball using counterfactual virtual simulation with deep learning. *International Journal of Computer Science in Sport*, 22(1):1–12, 2022.

P. M. Niedenthal, M. Mermillod, M. Maringer, and U. Hess. The simulation of smiles (sims) model: Embodied simulation and the meaning of facial expression. *Behavioral and brain sciences*, 33(6): 417–433, 2010.

L. M. Pereira, F. C. Santos, and A. B. Lopes. Ai modelling of counterfactual thinking for judicial reasoning and governance of law. 2023.

S. Rahimi, A. Moore, and P. Whigham. Towards intervention and counterfactual modelling in spatial agents: A simulation of constrained movement at the observational level. geocomputation 2019. 2019.

H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

N. J. Roese. The functional basis of counterfactual thinking. *Journal of personality and Social Psychology*, 66(5):805, 1994.

B. Sahoh, K. Haruehansapong, and M. Kliangkhlao. Causal artificial intelligence for high-stakes decisions: The design and development of a causal machine learning model. *IEEE Access*, 10: 24327–24339, 2022.

I. Saxena, G. Usha, N. Vinoth, S. Veena, and M. Nancy. The future of artificial intelligence in digital forensics: A revolutionary approach. In *Artificial Intelligence and Blockchain in Digital Forensics*, pages 133–151. River Publishers, 2023.

D. Schofield. Animating evidence: computer game technology in the courtroom. *Journal of Information, Law and Technology*, 1:1–21, 2009.

B. Shneiderman. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3):109–124, 2020.

F. L. Shults, W. J. Wildman, and V. Dignum. The ethics of computer modeling and simulation. In *2018 Winter simulation conference (WSC)*, pages 4069–4083. IEEE, 2018.

R. Singh, W. Wu, G. Wang, and M. K. Kalra. Artificial intelligence in image reconstruction: the change is here. *Physica Medica*, 79:113–125, 2020.

B. Smyth and M. T. Keane. A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In *International Conference on Case-Based Reasoning*, pages 18–32. Springer, 2022.

G. Solmazer, D. Azık, G. Fındık, Y. Üzümcüoğlu, Ö. Ersan, B. Kaçan, T. Özkan, T. Lajunen, B. Öz, A. Pashkevich, et al. Cross-cultural differences in pedestrian behaviors in relation to values: A comparison of five countries. *Accident Analysis & Prevention*, 138:105459, 2020.

S. Suo, S. Regalado, S. Casas, and R. Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10400–10409, 2021.

S. Surveswaran and L. Deshpande. A glimpse into the future: Ai, digital humans, and the metaverse–opportunities and challenges for life sciences in immersive ecologies. *AI in Clinical Medicine: A Practical Guide for Healthcare Professionals*, pages 521–527, 2023.

P. Suryavamsi and A. Arockia Selvakumar. Iot controlled mobile robot for home security and surveillance. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, pages 431–438. Springer, 2019.

Z. Tavares, J. Koppel, X. Zhang, R. Das, and A. Solar-Lezama. A language for counterfactual generative models. In *International Conference on Machine Learning*, pages 10173–10182. PMLR, 2021.

K. P. Tobia. How people judge what is reasonable. *Alabama Law Review*, 70:293–359, 2018.

A. Tolk, J. E. Lane, F. L. Shults, and W. J. Wildman. Panel on ethical constraints on validation, verification, and application of simulation. In *2021 Winter Simulation Conference (WSC)*, pages 1–15. IEEE, 2021.

M. Virgolin and S. Fracaros. On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence*, 316:103840, 2023.

J. Von Kügelgen, A. Mohamed, and S. Beckers. Backtracking counterfactuals. In *Conference on Causal Learning and Reasoning*, pages 177–196. PMLR, 2023.

K. D. Williams and A. Jones. Trial strategy and tactics. *Psychology and law: An empirical perspective*, pages 276–321, 2005.

S. A. Wu and T. Gerstenberg. If not me, then who? Responsibility and replacement. *Cognition*, 2023.

Y. Yacoby, B. Green, C. L. Griffin Jr, and F. Doshi-Velez. “if it didn’t happen, why would i change my decision?”: How judges respond to counterfactual explanations for the public safety assessment. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 219–230, 2022.

L. Yilmaz and B. Liu. Model credibility revisited: Concepts and considerations for appropriate trust. *Journal of Simulation*, 16(3):312–325, 2022.

H. Yuan and R. C. Veltkamp. Presim: A 3d photo-realistic environment simulator for visual ai. *IEEE Robotics and Automation Letters*, 6(2):2501–2508, 2021.

C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

Z. Zhang, Z. Yang, C. Ma, L. Luo, A. Huth, E. Vouga, and Q. Huang. Deep generative modeling for scene synthesis via hybrid representations. *ACM Transactions on Graphics (TOG)*, 39(2):1–21, 2020.