When AI meets Counterfactuals:
The Ethical Implications of Counterfactual World Simulation Models

Lara Kirfel
Max Planck Institute for Human Development

Rob MacCoun
Law School, Stanford University

Thomas Icard
Department of Philosophy, Stanford University

Tobias Gerstenberg
Department of Psychology, Stanford University

Abstract

This paper examines the transformative potential of AI embedded with *Counterfactual World Simulation Models* (CWSMs). A CWSM uses multimodal evidence, such as the CCTV footage of a road accident, to build a high-fidelity 3D reconstruction of what happened. It can answer causal questions, such as whether the accident happened because the driver was speeding, by simulating what would have happened in relevant counterfactual situations. We sketch 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 AI-powered 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. Sketching a new genre of AI, this paper seeks to illuminate the path forward for responsible and effective use of CWSMs.

Keywords: Counterfactual AI, Generative AI; AI Ethics; Counterfactual Simulation; Responsible AI; AI in the Law.

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 Artificial Intelligence (AI), this is about to change. Multi-modal generative AI models vastly expand the possibility of generating and interacting with evidence. Thanks to recent developments, such as scene simulation models in autonomous driving (Wayve, 2024), we can now reconstruct realistic simulations of what happened. Based on the CCTV footage, and a world model of car dynamics, street environments, and pedestrian behavior, such AI-powered simulation models are able to build a generative model of what happened and render a dynamic 3D simulation of how the crash came to pass (Gupta, Sharma, & Johri, 2020; Jadhav, Sankhla, & Kumar, 2020). In the near future, such models will not only be able to reconstruct what happened, but also run counterfactuals simulations of how things could have played out differently (see Li, Tian, Jiao, Chen, & Jiang, 2024; Tavares, Koppel, Zhang, Das, & Solar-Lezama, 2021; S. A. Wu et al., 2024; Z. Wu et al., 2023; L. Zhang et al., 2024). 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 help evaluating questions of responsibility and liability.

It's a long-standing goal in generative AI to develop dynamic and accurate 3D environments, allowing for dynamic simulations consisting of objects, spaces, and agents (A. Hu et al., 2023; Kaur, Singh, & Banerjee, 2023). Contemporary Multimodal Generative Models can perceive and integrate multi-modal inputs and perform counterfactual over reasoning of its content (Huang et al., 2023; Mondorf & Plank, 2024; Z. Wu et al., 2023). We will subsume all AI models that integrate diverse modalities into a world model and perform counterfactual reasoning over its objects under the term "Counterfactual World Simulation Models" (CWSMs; see Figure 2). Many of the building blocks of for CWSMs are already in place (see Figure 1). CWSMs represent an evolution from traditional image-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 (Brodeur et al., 2017; Clarke et al., 2022; Gan et al., 2020; Ivanovic, Schmerling, Leung, & Pavone, 2018; Li et al., 2018; Z. Zhang et al., 2020). While world simulation models can be used to predict what will happen next (Cui et al., 2020; Sahoh, Haruehansapong, & Kliangkhlao, 2022), and to infer what happened in the past, *counterfactual* world simulation models can also simulate counterfactuals scenarios of how things could have played out differently (Feder, Oved, Shalit, & Reichart, 2021; Gerstenberg & Stephan, 2021; Tavares et al., 2021; Vallverdú, 2024; Z. Wu et al., 2023).

Because counterfactual considerations about what would have happened are common practice in legal analysis and argumentation (Saxena, Usha, Vinoth, Veena, & Nancy, 2023), the application of generative AI in this domain could radically alter the landscape of legal

proceedings (Alarie, Niblett, & Yoon, 2018; Atkinson, Bench-Capon, & Bollegala, 2020). For example, we may ask whether the 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? 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 start out by describing what CWSMs are and elucidating the anticipated modes of interaction between these models and their users, as well as the ethical challenges that come with the capabilities they introduce. Subsequently, we explore several prospective applications of these models within the legal sphere, such as their role in generating evidence by legal fact-finders, and presenting such evidence in court. While the traffic accident will serve as our running example in this paper, AI-powered CWSMs can be applied in a variety of domains including sports analysis (e.g. "Would the game have developed differently if the player had been positioned differently?"), crowd and traffic control for event and emergency management or infrastructure decisions ("Would the crowd congestion have been reduced if the city had added additional gates for the concert?"), or the investigation of possible human rights violations ("Would the loss of civilian lives have been prevented if the military had taken alternative actions?", Weizman, 2017).

Counterfactual world simulation models

Simulations enable us to create digital replicas, 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 (Kapelyukh, Vosylius, & Johns, 2023; Yuan & Veltkamp, 2021). More recent generative AI models can handle multi-modal data. For example, 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 based on minimal input (Gozalo-Brizuela & Garrido-Merchán, 2023; C. Zhang, Zhang, Zhang, & Kweon, 2023). In the future, AI will be creating entire movies using text-to-video prompts (Ali, DiPaola, Williams, Ravi, & Breazeal, 2023). Open AI's text-to-video model Sora generates realistic and imaginative scenes based on textual instructions (Liu et al., 2024). As of now, the combination of Computer Vision and Natural Language Processing allows for the generation and manipulation of videos based on textual descriptions. Deep Generative Networks can artificially simulate a crime scene from textual input (Ashfaq, Jhanjhi, Khan, & Das, 2023), which can be used for crime and human rights investigation (Jacob, Thomas, & Savithri, 2024; Weizman, 2017). "What if, what now"

Figure 1. **Scene reconstruction models of the UK AI startup "Wayve" (Wayve, 2024)** A) By a video prompt (top left frame), the model is prompted to imagine what happens next. It generates an entirely new video clip where it imagines that the car drives forward after 2, 4 and 6 seconds. B) The model LINGO-2 can explain and respond to questions about the scene. In addition, the model can also provide a continuous driving commentary of its motion planning decisions, explaining its actions. C) A reconstruction of a scene with a pedestrian and D) that same scene in which the pedestrian is counterfactually removed.

GPT is a Contingency Scenario Planner using Generative AI for developing supply chain and event contingency scenarios (*what-if-Free Alternative History Simulator — yeschat.ai*, n.d.). Similarly, the "What-If game" is an AI-hosted simulation experience designed for exploring alternative historical scenarios (Finkenstadt, Eapen, Sotiriadis, & Guinto, 2023).

Integrating tools from causal inference with generative AI allows users to pose counterfactual queries and expand the powers of the simulation models even further: In addition to merely reconstructing evidence, generative AI is used for inquiring about what would have happened if certain factors in a situation would have been different (Feder et al., 2021; Tavares et al., 2021). Current multi-modal foundation models demonstrate counterfactual reasoning capabilities over textual data (Bhattacharjee, Moraffah, Garland, & Liu, 2024), images (L. Zhang et al., 2024) as well as video sequences (Mao, Yang, Zhang, Goodman, & Wu, 2022; T.-L. Wu et al., 2023). AI-assisted counterfactual generative simulation models provide powerful imagination machines that enable their users to entertain and depict counterfactual "what-if" scenarios, for example in autonomous driving simulations(cf. McDuff et al., 2022; Wang et al., 2024; Wayve, 2024). The AI startup "Wayve" has released a series

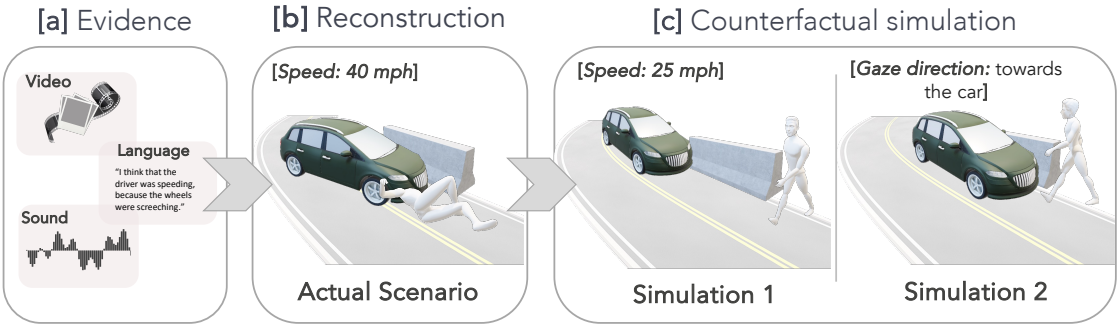

Figure 2. **Illustration of a Counterfactual World Simulation Model (CWSM) applied to a traffic accident.** The CWSM uses multi-modal evidence [a] to reconstruct what happened [b]. It can then simulate different counterfactual scenarios [c], to answer causal questions. For example, in Simulation 1, it 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 would have been avoided.

of models that enable photorealistic reconstructions in static and dynamic scenes. Their Generative AI world model GAIA-1 generates realistic driving videos from text and video inputs (Figure 1a) (A. Hu et al., 2023). LINGO-2 (Marcu et al., 2023) provides answers to questions about a driving scene and can explain factors that affect its driving behavior in natural language (Figure 1b). And PRISM-1 (Wang et al., 2024) reconstructs scene in space and time from video data (Figure 1c) and reasons about counterfactual changes to this scene (Figure 1c).

Counterfactual World Simulation models operate in a three-stage process, commencing with the collection and assimilation of multimodal evidence ("Evidence", Figure 2a), which could encompass various forms of data including images, videos, sounds, and textual information. With this evidence as the foundation, the model then proceeds to the reconstruction phase ("Reconstruction", Figure 2b), where it generates detailed and coherent representations of the observed events (Baltrušaitis, Ahuja, & Morency, 2018; Johnson, Alahi, & Fei-Fei, 2016). Following the accurate reconstruction of events, the counterfactual simulation stage ("Counterfactual Simulation", Figure 2c) (Tavares et al., 2021) allows for the exploration of various hypothetical situations and outcomes, offering insights into the myriad possibilities and consequences that could have arisen under different circumstances.

Evidence

What happened at the accident scene? In a first step to answer this question, an AI system will need to integrate multi-modal pieces of evidence from the scene, weaving together strands of information from diverse sources to build an accurate reconstruction of the events ("Evidence", Figure 2a). 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, Wu, Wang, & Kalra, 2020).

Reconstruction

Based on limited footage from CCTV and other pieces evidence, AI can then analyze the 2D images and videos to extrapolate depth information and reconstruct 3D scenes ("Reconstruction", Figure 2b). Embedded with a world model, the system can generate a simulation of what has happened, and thereby infer, for example, the car's velocity at the scene (Ashfaq et al., 2023). As a consequence, the model might reveal that the driver was speeding. But how do we know that the AI is right? If the simulation model will later help legal-fact finders draw inferences about liability and fault, users must trust it to display a high-fidelity (i.e., a high level of physical realism that closely matches what actually happened; Jacovi, Marasović, Miller, & Goldberg, 2021; Yilmaz & Liu, 2022) (Jacovi et al., 2021; Yilmaz & Liu, 2022).

Validating the reconstruction accuracy. A generative simulation model that creates a realistic simulation from limited visual input will need to be validated against the ground truth (Tolk, Lane, Shults, & Wildman, 2021). Cross-verification from a variety of data sources can be used to validate the AI's output. Real-world data from traffic sensors, CCTV, 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. An additional strategy to ensure a model's validity is to test its predictions on test 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 prediction can then be compared against the video footage of what actually happened. Future reliance on scene simulation models in legal proceedings demands appropriate trust based on credible results (Yilmaz & Liu, 2022).

Counterfactual simulation

Synthesizing and reconstructing evidence, the AI 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 models can then be used to simulate counterfactual scenarios (Tavares et al., 2021). Now that we know that the driver has been 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 AI 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 generative AI needs to interpret and instantiate this counterfactual scenario (Lassiter, 2017a, 2017b). 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, 2022; Gerstenberg & Stephan, 2021; Goodman, 1947; Von Kügelgen, Mohamed, & Beckers, 2023)?

In order to answer such counterfactual queries, such as what role the car's speeding played in the accident, 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, von Kügelgen, Schölkopf, & Valera, 2020; Virgolin & 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 case of a lack of distinctive feature instantiations (e.g. when no clear speed limit is available), the model 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 (e.g. average driving speed) and prescriptive (e.g. generally allowed or recommended driving speed) feature or behavior (cf. Bear & Knobe, 2017). For example, the simulated driving speed for a car on an interstate highway will be an average between the speed limit, 65mph, and the average speed that people actually drive, 75mph – so 70mph. Depending on the context, the statistical average and the normative value can be weighted differently. The model might also consider multiple scenarios that fit the query, perhaps generating a distribution of outcomes based on different interpretations of "slower" or "driving below the 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.

In some scenarios, the outcome may have been overdetermined by several causes that would have been individually sufficient. Such situations of redundant causation were traditionally considered problem cases for counterfactual theories of causation (Lewis, 2000). For example, consider a situation in which an unlucky pedestrian is struck in the head because of a collision between two cars that approached from opposite directions. Intuitively, we would want to say that both cars caused the pedestrian's injuries even though the pedestrian would still have been injured even if just one of the two cars had struck him. To deal with such situations of overdetermination, counterfactual theories developed an extended definition of counterfactual dependence to capture causation (e.g., Chockler & Halpern, 2004; Halpern & Pearl, 2005). Accordingly, an event can count as a cause even if it made no difference in the actual situation, as long as it would have made a difference in a relevant counterfactual contingency. In this case, each car is a cause because it would have made a difference if the other one hadn't been there. These theories define general principles that constrain what kinds of counterfactual contingencies may be considered to assess causation. Another approach to accommodate the intuition that both cars caused the pedestrian's injuries is to think about the causal connection at different levels of granularity. For example, while neither of the two cars individually made a difference to *whether* the pedestrian was injured, they each make a difference to *how* the injuries came about (Gerstenberg, 2024; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Lewis, 2000).

To capture whether-causation, one considers the absence of the cause (what would have happened if the car hadn't been there). To capture how-causation, one considers a small perturbation to the cause (what would have happened if one of the cars had driven a little faster or slower) and checks whether that would have made a difference to the outcome event, finely construed (i.e., the exact timing and manner of the outcome).

One of the key advantages of using a counterfactual criterion of causation over other choices such as defining causation in terms of processes (e.g., Salmon, 1984; Wolff, 2007) or statistical regularities (e.g., Andreas & Günther, 2024; Baumgartner & Falk, 2023), is that counterfactuals flexibly apply to causation by omission, too. For example, we may want to say that a driver's failure to break on a red light was responsible for the accident (Gerstenberg & Stephan, 2021; Henne, Pinillos, & De Brigard, 2017). Normative expectations play a particularly important role when determining which omissions are deemed relevant for an outcome (McGrath, 2005). Generally, we tend to cite omissions as causes when we expected the person to act and they failed to do so.

Constraints on counterfactual simulations. We might be interested not only in the causal role of the driver, but also in whether the pedestrian had any fault in the incident. Could the pedestrian have behaved differently such that the accident would have been avoided? On the ethical level, the general question arises as to what simulations should be admissible (Fazelpour, 2021). AI-assisted simulation models will be able to generate any 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 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?

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 the simulation of 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 & Goodwin, 2016). 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, e.g. a social group, the model must infer the primary characteristics and fill in all information that is underspecified. Text-to-Image AI might hence render neutral genderless language into visually exclusive and biased representations (Bianchi et al., 2023; Luccioni, Akiki, Mitchell, & Jernite, 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.

How to de-bias a simulation model? Implementing user-side guardrails in counterfactual simulation models would involve placing restrictions on the types of simulations that users can request or initiate, to ensure responsible and ethical use. Algorithmic fairness audits can help identify and rectify biases (De Schutter & 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? It is worth noting that the use of social categories such as race and gender for counterfactual manipulation has been critiqued in the context of algorithmic fairness (L. Hu & Kohler-Hausmann, 2020; Kasirzadeh & Smart, 2021). Current methodologies, it has been argued, fail to acknowledge the socially constructed nature of these variables. They treat social categories such as race and gender as a simple, manipulable attribute rather than accounting for the structural and systemic factors that influence them (L. Hu & Kohler-Hausmann, 2020; Kohler-Hausmann, 2018). Counterfactual causal models typically assume that one can isolate and modify one causal pathway without affecting others, but this is not possible with non-modular, interdependent social phenomena (L. Hu & Kohler-Hausmann, 2020). L. Hu and Kohler-Hausmann (2020) suggest instead to focus on constitutive relations rather than causal relations, that is, relationships and features that define and constitute a social category. CWSMs could integrate such a constructivist approach to social variables such as gender by dynamically adjusting factors that go beyond binary biological markers including social and psychological traits, and highlight a network of factors that constitute a social variable. There remain formidable challenges on this front.

While high representational fidelity of an agent might be desirable at the stage of reconstructing the evidence, it can lead to bias at the stage of simulating counterfactuals. In the initial phase of reconstructing the evidence—the simulation of the actual scenarios—it's crucial for the AI to have high representational fidelity. However, when the model includes additional attributes of a person compared to a generic person in a model representation, this might lead to bias at the stage of counterfactual simulation. For example, it should generate a simulation that accurately depicts a teenager walking across the street. However, in order to simulate a scenario to answer the question, "Would the accident have been avoided if the pedestrian had behaved differently?" the AI, because of the high representational fidelity from the reconstruction phase, might automatically associate distracted behavior with a certain age group and rely on the stereotype that young individuals tend to be distracted and careless about their surroundings. This could lead the AI 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. A model's physics engine can represent the body of objects without rendering the shape (i.e. the textures including skin color, clothes, hair, etc.), but should represent the height and other physical properties (e.g. how heavy the person is) because these factors are critical for physical simulations. We 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 or gendered cues. In cases where macro behaviors like pedestrian paths are in focus, simulation can represent agents as abstract symbols or icons that convey their movements and interactions.

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 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 proportional with regards to varying parameters like walking speed, reaction time and awareness of surroundings, and culturally 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). 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 & Janssen, 2002; Suo, Regalado, Casas, & Urtasun, 2021).

While ethical standards need to be in place in order to preserve the integrity of simulated agents, interacting with counterfactual simulation models can also have a psychological impact on its users. The vivid realism of these simulations can evoke intense emotional reactions, especially when the simulation involves traumatizing events, making simulated scenarios feel almost as impactful as lived experiences (Coricelli & Rustichini, 2010). This is particularly potent when the simulation resonates personally with the user, potentially triggering memories or emotions tied to the experienced real-life events. In certain cases, consent may become required when simulations involve generating images of people whose likeness is used, sensitive behaviors, or private spaces.

Interacting with a counterfactual world simulation model

Design and accessibility of a counterfactual simulation model system plays a pivotal role in shaping user interactions through the kinds of counterfactual scenarios they will probe. As inputs are manipulated, CWSMs offer immediate 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.

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., Goodman, 1947; Lassiter, 2017a, 2017b). On the other hand, coded prompts or parameter-setting interfaces offer more precise control over the counterfactual conditions but may require a deeper understanding of the model's workings (Bove, Lesot, Tijus, & Detyniecki, 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 (GUI) 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).

Model output

After having established which counterfactual simulations an AI 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 generative AI model 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, Green, Griffin Jr, & Doshi-Velez, 2022)? The fact that these systems produce multiple simulations with varying degrees of probability 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, Zheng, & Lallie, 2010). How can we improve their presentation to facilitate a more nuanced understanding? We suggest the sequential presentation of a subset of simulations, in randomized order, 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 flip the outcome. In addition, counterfactual simulations should be plausible and feasible, they should modify features that make sense to users (e.g., driver's behavior vs. changes in traffic laws or structure of the street environment) and they should 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 & 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, Hulsebos, Gale, Haas, & Demiralp, 2021; Shneiderman, 2020).

Applications of counterfactual world simulation models

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, Santos, & Lopes, 2023). While counterfactual reasoning is not new to legal practice, the automation and visual representation provided by AI-powered simulations introduces distinct ethical and procedural challenges.

CWSMs as tools for aiding legal fact finders

Simulations can provide a powerful and persuasive tool in legal court cases. AI-generated simulations can provide a level of detail that would not be possible to achieve through verbal or written descriptions alone. For example, a simulation of a car accident might include precise data on vehicle speeds, pedestrian movements, and environmental conditions, all visualized in a realistic manner. This granularity can highlight minor details, such as the exact timing of a pedestrian's decision to step into the street or the specific moment a driver applied the brakes. As a consequence, they need to meet several restrictions and regulations to ensure fair and ethical use (Schofield, 2009). When it comes to the

question of, if and how an agent could have behaved differently in a certain scenario, not all simulations that are physically possible and ethically sound should be allowed to be introduced 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 introduction 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 those that do not make a difference to the outcome (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. However, we caution that the simulation of agent behavior displayed in legal proceedings should be constrained to contextually relevant behavior unless parties can provide good reasons for doing otherwise. Simulations should not be used to divert attention away from the main issues of a case and focus on actions that are directly related or relevant to the actual outcome in question, e.g. the accident.

Simulations should not unjustly portray the agent as at fault or as acting negligently or recklessly without appropriate supporting evidence (Shults, Wildman, & Dignum, 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 the emotional impact that such events would normally elicit, and be used as a tool for desensitization (Williams & 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. 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 tailored 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.

CWSMs as tools for presenting evidence in court

At the end of a counterfactual simulation cycle stands the potential to use its results as expert testimony in legal proceedings. The persuasive powers of realistic simulations as evidence in court, however, pose challenges for accuracy and procedural fairness because of the potential for an uncritical belief in the presented material (Clifford & 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. However, the realistic rendering of components of the virtual model may possibly lull the viewer into a "seeing-is-believing" attitude (Etienne, 2021; Ma et al., 2010).

We advocate for clear guidelines for using AI simulations as evidence by proposing limitations on the level of detail for simulation objects, agents, and environments when presented in court. Simulations should display high physical fidelity of the environment 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, Mermillod, Maringer, & Hess, 2010). This will minimize over-persuasion powers (Ma et al., 2010) and respect agents' dignity, or, in certain cases, their anonymity or right to privacy. Facial expressions can affect observers' inferences about agents' hypothetical emotional and intentional states and influences observers' judgments about the agent. For example, the model could inaccurately portray the agent as angry or distressed. Limiting emotional expressions can mitigate this. Similarly, avoiding misleading emotional portrayals of victims or witnesses, and not showing individuals' expressions in potentially humiliating situations, preserves their dignity in court. In general, features should only be made explicit when they can be shown to have some causal relevance for the outcome and are variables of interest, for example, age for simulating walking speed, height for accessibility, etc. This also applies to facial expressions: A person's expression will only be modeled when it's causally relevant for the episode. For example, if an agent's surprise influenced another agent's actions in the situation, such as causing them to pause or change their decision, then modeling the surprise expression is necessary. Otherwise, facial expressions do not need to be displayed in a simulation.

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 allow 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. The specific implementation and procedural details may vary depending on the legal jurisdiction, nature of the case, and domain-specific requirements.

Conclusion: Counterfactual AI – 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 resulting insights and knowledge. We have outlined some of the normative, practical and legal challenges, and offered some proposals for how we might 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. We identify some of the ethical issues that will need to be addressed in a more normatively grounded framework. Our paper provides an overview and guide to some of the most pressing and distinctive issues for this new kind of AI.

Acknowledgments

We thank David Danks and Johannes Himmelreich for helpful discussions. This work was supported by a grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

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

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

Andreas, H., & Günther, M. (2024). A regularity theory of causation. *Pacific Philosophical Quarterly*, *105*(1), 2–32.

Ashfaq, F., Jhanjhi, N. Z., Khan, N. A., & Das, S. R. (2023). Synthetic crime scene generation using deep generative networks. In *International conference on mathematical modeling and computational science* (pp. 513–523).

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

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

Baumgartner, M., & Falk, C. (2023). Boolean difference-making: a modern regularity theory of causation. *The British Journal for the Philosophy of Science*.

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

Bhattacharjee, A., Moraffah, R., Garland, J., & Liu, H. (2024). Zero-shot llm-guided counterfactual generation for text. *arXiv preprint arXiv:2405.04793*.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., ... Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 acm conference on fairness, accountability, and transparency* (pp. 1493–1504).

Bove, C., Lesot, M.-J., Tijus, C. A., & Detyniecki, M. (2023). 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* (pp. 188–203).

Brodeur, S., Perez, E., Anand, A., Golemo, F., Celotti, L., Strub, F., ... Courville, A. (2017). Home: A household multimodal environment. *arXiv preprint arXiv:1711.11017*.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*(1), 93–115.

Clarke, S., Heravi, N., Rau, M., Gao, R., Wu, J., James, D., & Bohg, J. (2022). Diffimpact: Differentiable rendering and identification of impact sounds. In *Conference on robot learning* (pp. 662–673).

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

Coricelli, G., & Rustichini, A. (2010). Counterfactual thinking and emotions: regret and envy learning. *Philosophical Transactions of the Royal society B: Biological sciences*, *365*(1538), 241–247.

Cui, P., Shen, Z., Li, S., Yao, L., Li, Y., Chu, Z., & Gao, J. (2020). Causal inference meets machine learning. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* (pp. 3527–3528).

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

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

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

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

Finkenstadt, D. J., Eapen, T. T., Sotiriadis, J., & Guinto, P. (2023, Nov). Retrieved from `https://hbr.org/2023/11/use-genai-to-improve-scenario-planning`

Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., ... others (2020). Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*.

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

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

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

Gerstenberg, T. (2024). Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(6), 936–975.

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, 104842.

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

Goodman, N. (1947). The problem of counterfactual conditionals. *The Journal of Philosophy*, *44*(5), 113–128.

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

Gupta, S., Sharma, M. V., & Johri, P. (2020). Artificial intelligence in forensic science.

International Research Journal of Engineering and Technology, *7*(5), 7181–7184.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887.

Henne, P., Pinillos, Á., & De Brigard, F. (2017). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, *95*(2), 270–283.

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

Hu, L., & Kohler-Hausmann, I. (2020). What's sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770*.

Huang, Y., Hong, R., Zhang, H., Shao, W., Yang, Z., Yu, D., ... Song, L. (2023). Clomo: Counterfactual logical modification with large language models. *arXiv preprint arXiv:2311.17438*.

Ivanovic, B., Schmerling, E., Leung, K., & Pavone, M. (2018). Generative modeling of multimodal multi-human behavior. In *2018 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 3088–3095).

Jacob, L., Thomas, K., & Savithri, M. (2024). Ai in forensics: A data analytics perspective. In *Artificial intelligence for cyber defense and smart policing* (pp. 41–60). Chapman and Hall/CRC.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). 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* (pp. 624–635).

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

Jager, W., & Janssen, M. (2002). The need for and development of behaviourally realistic agents. In *International workshop on multi-agent systems and agent-based simulation* (pp. 36–49).

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). 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* (pp. 694–711).

Jusiega, V. (2022). *Designing a user interface for counterfactual simulations of adaptive treatment strategies* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

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

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., & Valera, I. (2020). Towards causal algorithmic recourse. In *International workshop on extending explainable ai beyond deep models and classifiers* (pp. 139–166).

Kasirzadeh, A., & Smart, A. (2021). The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 228–236).

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

Kohler-Hausmann, I. (2018). Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, *113*, 1163.

Lassiter, D. (2017a). Complex antecedents and probabilities in causal counterfactuals. In *Proceedings of the 21st amsterdam colloquium* (pp. 45–54).

Lassiter, D. (2017b). Probabilistic language in indicative and counterfactual conditionals. In *Semantics and linguistic theory* (Vol. 27, pp. 525–546).

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

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197.

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

Li, Y., Tian, W., Jiao, Y., Chen, J., & Jiang, Y.-G. (2024). Eyes can deceive: Benchmarking counterfactual reasoning abilities of multi-modal large language models. *arXiv preprint arXiv:2404.12966*.

Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., ... others (2024). Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*.

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

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

Mao, J., Yang, X., Zhang, X., Goodman, N., & Wu, J. (2022). Clevrer-humans: Describing physical and causal events the human way. *Advances in Neural Information Processing Systems*, *35*, 7755–7768.

Marcu, A.-M., Chen, L., Hünermann, J., Karnsund, A., Hanotte, B., Chidananda, P., ... others (2023). Lingoqa: Video question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*.

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

McDuff, D., Song, Y., Lee, J., Vineet, V., Vemprala, S., Gyde, N. A., ... Kapoor, A. (2022). Causalcity: Complex simulations with agency for causal discovery and reasoning. In *Conference on causal learning and reasoning* (pp. 559–575).

McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies*, *123*(1), 125–148.

Mondorf, P., & Plank, B. (2024). Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*.

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

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

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

Sahoh, B., Haruehansapong, K., & Kliangkhlao, M. (2022). Causal artificial intelligence

for high-stakes decisions: The design and development of a causal machine learning model. *IEEE Access*, *10*, 24327–24339.

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

Saxena, I., Usha, G., Vinoth, N., Veena, S., & Nancy, M. (2023). The future of artificial intelligence in digital forensics: A revolutionary approach. In *Artificial intelligence and blockchain in digital forensics* (pp. 133–151). River Publishers.

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

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

Shults, F. L., Wildman, W. J., & Dignum, V. (2018). The ethics of computer modeling and simulation. In *2018 winter simulation conference (wsc)* (pp. 4069–4083).

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

Smyth, B., & Keane, M. T. (2022). A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In *International conference on case-based reasoning* (pp. 18–32).

Solmazer, G., Azık, D., Fındık, G., Üzümcüoğlu, Y., Ersan, Ö., Kaçan, B., ... others (2020). Cross-cultural differences in pedestrian behaviors in relation to values: A comparison of five countries. *Accident Analysis & Prevention*, *138*, 105459.

Suo, S., Regalado, S., Casas, S., & Urtasun, R. (2021). Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10400–10409).

Tavares, Z., Koppel, J., Zhang, X., Das, R., & Solar-Lezama, A. (2021). A language for counterfactual generative models. In *International conference on machine learning* (pp. 10173–10182).

Tolk, A., Lane, J. E., Shults, F. L., & Wildman, W. J. (2021). Panel on ethical constraints on validation, verification, and application of simulation. In *2021 winter simulation conference (wsc)* (pp. 1–15).

Vallverdú, J. (2024). Counterfactual thinking for machines. In *Causality for artificial intelligence: From a philosophical perspective* (pp. 63–76). Springer.

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

Von Kügelgen, J., Mohamed, A., & Beckers, S. (2023). Backtracking counterfactuals. In *Conference on causal learning and reasoning* (pp. 177–196).

Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., ... Alvarez, J. M. (2024). Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*.

Wayve. (2024). *Introducing prism-1: Photorealistic reconstruction in static and dynamic scenes.* `https://wayve.ai/thinking/prism-1/`. (Accessed: 2024-06-28)

Weizman, E. (2017). *Forensic architecture: Violence at the threshold of detectability*. Princeton University Press.

what-if-Free Alternative History Simulator — yeschat.ai. (n.d.). `https://www.yeschat.ai/gpts-9t55QeIsx7e-what-if`. ([Accessed 13-07-2024])

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

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

Wu, S. A., Brockbank, E., Cha, H., Fränken, J.-P., Jin, E., Huang, Z., ... Gerstenberg, T. (2024). Whodunnit? inferring what happened from multimodal evidence. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Wu, T.-L., Dou, Z.-Y., Hu, Q., Hou, Y., Chandra, N. R., Freedman, M., ... Peng, N. (2023). Acquired: A dataset for answering counterfactual questions in real-life videos. *arXiv preprint arXiv:2311.01620*.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., ... Kim, Y. (2023). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Yacoby, Y., Green, B., Griffin Jr, C. L., & Doshi-Velez, F. (2022). "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* (Vol. 10, pp. 219–230).

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

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

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

Zhang, L., Zhai, X., Zhao, Z., Zong, Y., Wen, X., & Zhao, B. (2024). What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 21853–21862).

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