Anticipating the Risks and Benefits of Counterfactual World Simulation Models (Extended Abstract)

Lara Kirfel[1*], **Rob MacCoun**[2], **Thomas Icard**[2], **Tobias Gerstenberg**[2]

[1]Center for Humans and Machines
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany
[2]Stanford University
450 Jane Stanford Way, Stanford, CA 94305, United States

Abstract

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, Sharma, and Johri 2020; Jadhav, Sankhla, and Kumar 2020). Such models will not only be able to reconstruct what happened, but also run 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.

Here, we focus on a class of generative simulation models that we call "Counterfactual World Simulation Models" (CWSMs). 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, Haruehansapong, and Kliangkhlao 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).

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, Niblett, and Yoon 2018; Atkinson, Bench-Capon, and Bollegala 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. See the full paper version here: **https://osf.io/preprints/psyarxiv/9x8rv**

[*]Corresponding Author: Lara Kirfel, kirfel@mpib-berlin.mpg.de