*Proceedings of the 2023 Winter Simulation Conference*
*C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.*

# THE GROWTH OF GENERATIVE AI: HYPE, HARM, AND CONTROL

Timothy Clancy

Dialectic Simulations Consulting, LLC
2240 Mohawk Trail
Acworth, GA 30102, USA

Asmeret Naugle

Sandia National Laboratories
PO Box 5800 MS 1327
Albuquerque, NM 87185, USA

Ignacio J. Martinez-Moyano

Argonne National Laboratory
9700 South Cass Avenue
Lemont, IL 60439, USA

## ABSTRACT

The *hype-harm-control* model investigates the societal impact of generative artificial intelligence (AI), given its growth, alignment with societal values, and controls. This system dynamics model was used to simulate the impact of generative AI over a 10-year time horizon. As the generative AI grows, hype and use increase, leading to both societal benefit and societal harm. This analysis found that while the balance of hype and societal harm determines the controls put on AI development and use, early societal harm creates a strong incentive to implement societal controls that limit the growth of generative AI overall.

## 1 INTRODUCTION

Generative AI has taken the world by storm, and its ultimate impact will depend on how its uses evolve as technology advances. Novel applications like ChatGPT and Stable Diffusion, have captivated the world's attention by generating human-like conversation and innovative images. However, initial dangerous uses of AI, such as disinformation and deepfake generation, may point toward future AI-related problems. A key question arises: as AI grows, can humanity's capacity to regulate AI grow fast enough to prevent significant harm to society?

## 2 THE HYPE-HARM-CONTROL MODEL

The *Hype-Harm-Control* model uses system dynamics modeling to simulate the character of future AI growth and how this will impact the balance of AI's benefits and harms to society (see Figure 1). In the figure, a positive ("+") sign represents a positive causal link and means that, all else equal, the two variables will follow the same trajectory over time; increases (or decreases) in the variable at the beginning of the arrow will result in increases (or decreases) in the variable at the end of the arrow. A negative causal link ("–") means that, all else equal, the two variables will follow opposite trajectories over time. Variables inside of rectangles represent accumulations in the system and the colors used for both the rectangles and the connecting arrows are meant as an aid to identify the influences of the different variables in the model. The model simulates AI capability growth, which creates hype within the social psyche about AI. As hype grows the AI industry grows, further building AI capability. Hype also influences people to use new AI capabilities. Users further discuss the benefits of AI, boosting hype even more. As capabilities and the use of AI grows, people find ways to use these new tools to benefit society. But not all uses of AI are socially beneficial. Some bad actors will find nefarious uses of these new AI capabilities, causing harm to society. This harm will reduce the hype around AI, creating skepticism and reluctance. But society will try to counteract these negative consequences, putting more resources into controlling AI and pushing it to align with beneficial goals of society at large. This alignment is the key factor that will determine whether AI creates a net benefit to humanity.

This structure is similar to what we see with many new technological revolutions, but AI has one key difference from other examples: AI has been trained to write code and at some point, perhaps the very near future, it may be able to develop new AI and evolve itself without human interference. This potential for self-generation makes AI a truly unique technological innovation, and gives it potential to excel past any human expectations, with possible nefarious or benevolent directions. The key in determining whether the direction is good or evil lies in whether we can ensure AI alignment to society's values.

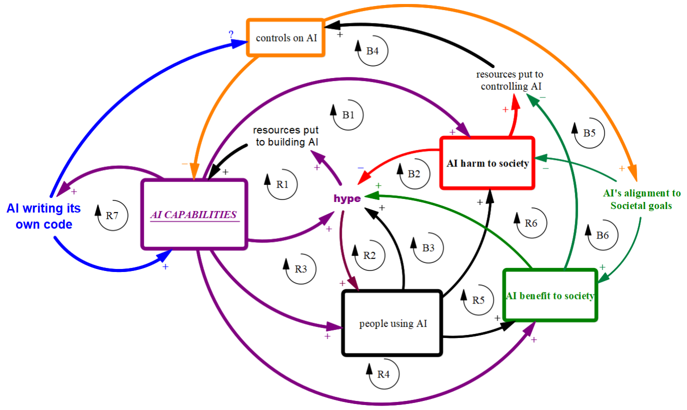

Figure 1: Hype-Harm-Control model.

## 3 RESULTS

We focus here on the key scenario of interest: the hype-leads-to-harm scenario. In this scenario, we introduce a sudden "spark" of industry interest in AI. Hype about AI increases, more people use new AI capabilities, and the industry, and thus capability, grows even more. We ran two scenarios with different levels of AI dis-alignment with society's values, 0.25 and 0.75, respectively (see results in Figure 2).

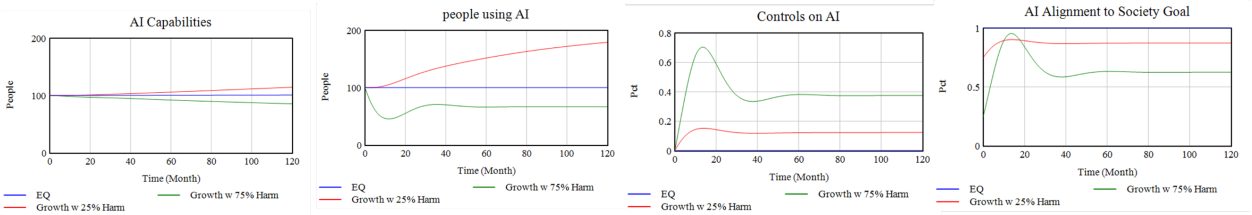

Figure 2: Hype-Harm-Control model results.

When dis-alignment is high, AI capabilities decrease, as hype diminishes and people stop using AI capabilities. In this case, a significant effort is made to control AI, leading to an initial boost in AI alignment, which falls again after the decline of the industry. The key insight here isn't just that technology can provoke harm, or that harm can provoke control response. Instead, it is the power and velocity of harm, and it's accumulation, that triggers relatively more powerful or weaker control responses. The slower harm accumulates, the less severe the control response will be. Several key findings can be explored in the Hype-Harm-Control model. First, there is always some harm and disruption within the hype cycles of new technology adoption. Second, the relationship between high hype fueling growth and the velocity at which harm accumulates determines the strength and severity of control reaction. Too high a harm, too early, creates a stronger societal and control backlash that limits overall growth. However some initial disruptive harm can be absorbed as controls result in an alignment adjustment, reducing the harm sufficiently to allow the hype cycle to continue and the new technology to become a permanent feature.