AI Horizons White Paper

Ethics and Society

Last Updated 2-19-2025

GenAI is likely to have broad and deep implications for the way we work, the things we create, the way we come to know things, and the relationships we form. This has created substantial anxiety with respect to GenAI's impact on our economy, our politics, and our culture(s). Many of these overlap with concerns that faculty have about the ways in which GenAI will change research and teaching at universities. In this white paper, we focus on four key areas of research that we think: (i) have the greatest potential for impact, (ii) align with existing expertise or interests at UR, and (iii) have significant potential for transdisciplinary collaboration. These are: (1) Safety and Alignment, (2) Work, (3) Trust, and (4) Creativity. We make two structural observations.

First, one of these research domains is unlike the others. While work, trust and creativity broadly examine the *impact* of GenAI broadly construed, safety and alignment are broadly about *creating* GenAIs with certain traits. In this respect, we can think of the "alignment" research theme as focused on questions that are prior to the deployment of a GenAI, while "work, "trust" and "creativity" research themes refer to questions that arise downstream of the deployment of a particular GenAI. Obviously, there is considerable overlap between upstream and downstream research as we try to implement solutions.

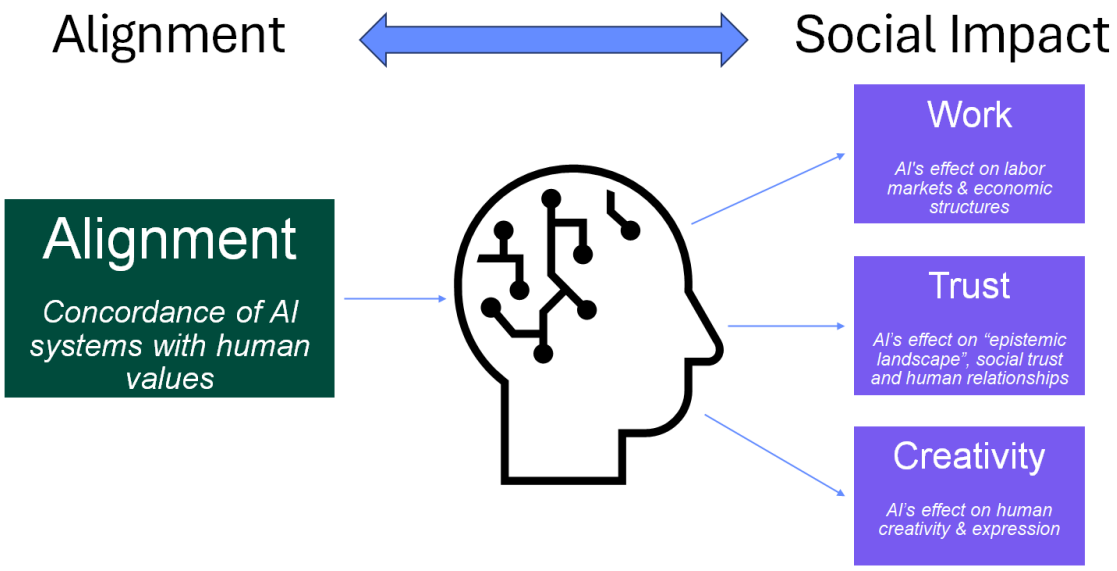

Figure 1: Conceptualizing the ethical space.

Second, the breadth of these domains - and the huge range of potential methodologies one could use to study them - requires us to compartmentalize the research questions into manageable research programs. To do this we can chunk research questions into three broad domains (see Figure 2).

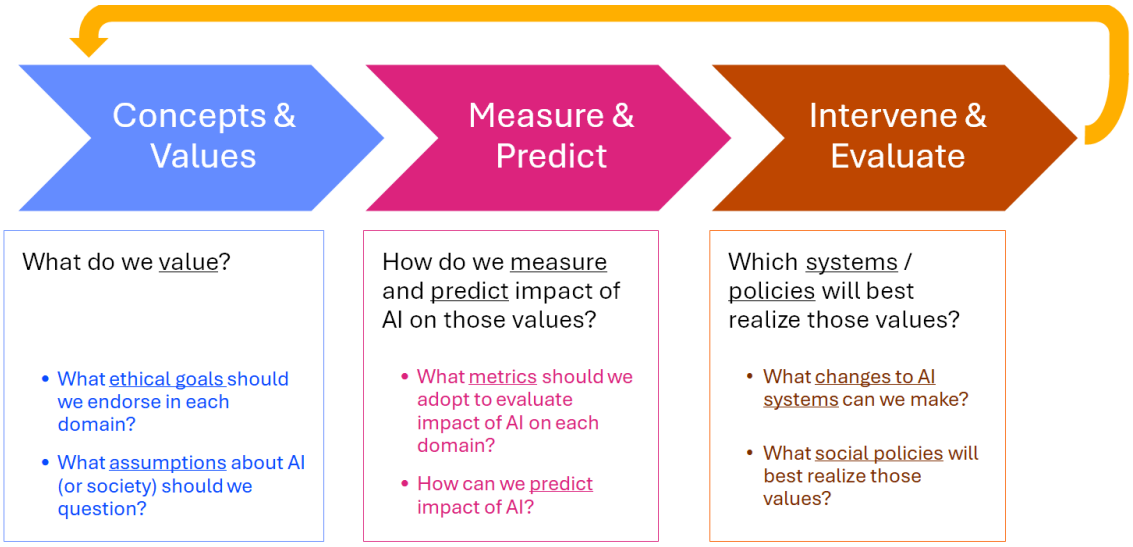

Figure 2: Broad research questions.

Defining concepts and values requires careful examination of how AI challenges and reshapes fundamental social constructs. Research questions in this domain might explore:

- How do we define meaningful human control in an era of increasingly autonomous systems?
- What constitutes genuine creativity when AI can generate sophisticated artistic outputs?
- How should we conceptualize privacy when AI systems can infer intimate details from seemingly innocuous data?
- What does fairness mean in algorithmic decision-making across different cultural contexts?
- How do we balance innovation with preservation of traditional practices and knowledge systems?

These definitional questions require integrating perspectives from philosophy, social science, computer science, and affected communities.

Measuring and predicting change demands robust methodologies for tracking AI's societal impact. Key research questions include:

- How can we measure changes in human agency and autonomy as AI systems become more prevalent?
- What metrics effectively capture shifts in creative processes and outputs?

- How do we assess changes in trust between humans and across human-AI interactions?
- What indicators best reveal transformations in labor markets and skill requirements? How can we track the evolution of social norms around AI use?

These questions require developing new measurement frameworks that combine quantitative and qualitative social science approaches, while acknowledging the challenge of capturing complex social phenomena.

Intervening and evaluating focuses on actionable strategies to shape AI's societal integration. UR researchers might ask:

- What policy mechanisms effectively promote beneficial AI development while mitigating risks?
- How can educational systems be reformed to prepare people for AI-augmented work environments?
- What design principles support AI systems that enhance rather than diminish human capabilities?
- How can we structure incentives to align commercial AI development with public interest?
- What governance frameworks enable democratic oversight while fostering innovation?

These questions emphasize practical approaches to realizing identified values, requiring careful consideration of technical feasibility, social acceptance, and implementation challenges across different contexts and scales.

I. Main Areas of Research

This white paper identifies four critical domains requiring sustained research attention as artificial intelligence becomes increasingly integrated into society. Safety and Alignment examines the technical and social mechanisms needed to ensure AI systems behave reliably and in accordance with human values. Work investigates how AI is reshaping labor markets, skill requirements, and organizational structures across industries. Trust explores the socio-technical dimensions of human-AI interaction, including transparency, accountability, and the development of appropriate reliance. Creativity analyzes AI's impact on human creative expression, artistic production, and cultural evolution. Whilst these domains are deeply interconnected, with developments in one area significantly influencing outcomes in others; they form distinct enough research areas that cohesive teams can be formed. Moreover, they each contain research questions that follow our model of (i) defining concepts and values, (ii) measuring and predicting change, and (iii) identifying ways to intervene to best realize our values. Finally, each domain involves both technical and social dimensions that must be considered holistically to develop effective solutions and governance frameworks.

A. Alignment

A1. Estimating the Risks of GenAI

A range of research has tried to survey the ethical risks posed by GenAI.[1] These main areas of concern have involved investigating both *unintended* risks and *malicious* misuse. For unintended risks, the main areas of concern include:

1. Bias: GenAI models trained on human-generated data could amplify and propagate existing biases around gender, race, politics, and other sensitive topics that are embedded in the training data (or imposed by AI developers). This involves a suite of research on "toxicity", "representational bias", and "content moderation" in large language models as well as in image and video generators.
2. Robustness: GenAI models can confabulate or make errors in subtle ways, and rarely offer users insights into the source of the model's judgements (although Google and perplexity.ai have made efforts to mitigate this in their models) This involves a suite of research on "grounding" (cf. Harnad 1990), i.e., about how models can develop a more robust understanding of the expressions they use. Moreover, GenAI models have been shown to exhibit other types of undesirable behavior, such as sycophancy and deception (Sharma et al. 2023, Park et al. 2024).
3. Transparency & Accountability: The complexity of GenAI models makes their decision-making processes opaque, limiting our ability to audit and hold them (or their users) accountable. Much of this research revolves around the so-called "responsibility gap", and the relative merits of different approaches to ensuring "explainability" or "interpretability" of decisions made using AI.

Most of the research on unintended risks revolves around identifying *metrics* of the different risks (i.e. bias, grounding etc.), *audit studies* that attempt to probe individual models for these risks, and *methods* for removing or ameliorating these risks in various models. **UR has some existing research in this space, mostly on vision models (Chenliang Xu).**

For malicious misuse, the main areas of concern are:

1. Deceptive Content Generation: Malicious actors could use GenAI to create highly convincing disinformation, fake news, phishing emails, and other deceptive content at scale. (see section "Trust" below)
2. Use for Crime/Harm: GenAI could be used to assist in committing crimes or other kinds of harmful actions. For instance, they could be used to assist in cyberattacks or in developing chemical or biological weapons (OpenAI 2024). This includes the use of GenAI in creating novel (and possibly undetectable) methods for doing harm, such as generating novel forms of malicious code, new chemical or biological weapons, or instructions on covering up major crimes.

Most of the research on malicious misuse has focused on "red-teaming" and adversarial attacks, where experts attempt to use prompt engineering techniques or external tools to "jailbreak"

content filters and elicit unsafe results. Extant countermeasures include “watermarking” GenAI output to make them traceable to their source, and the use of Reinforcement Learning from Human Feedback (RLHF) to prevent jailbreaks.

A2. Identifying and Measuring Human Values

In order to “align” GenAI with human values, we need to identify what those values are (or ought to be). Two problems make this task more difficult. First, there is disagreement about what makes for a just, fair or good decision across different peoples, places and periods of time. Often these disagreements are not just over which ethical values to prioritize, but over the epistemological and metaphysical assumptions that are embedded in different value systems. Second, we need to find a legitimate way of navigating that disagreement to come to collective decisions about which values to embed in AI. Different philosophical theories of justice suggest different ways of resolving disagreements. The utilitarian approach is the most straightforward: it recommends choosing the algorithm that maximizes utility in the aggregate.[2,3] This approach is often considered unsatisfying as a way of aggregating disagreements because it ignores the rights of minority value sets. One of the most influential alternative theories are “veil of ignorance” approaches,[4,5] that asks individuals to design an AI system behind a “veil of ignorance”, without knowledge of their individual values or social position. Importantly, owing to our strengths in political science, UR **has a comparative advantage in voting theory, preference elicitation, and formal theories of justice.**

A3. Embedding Human Values in GenAI

A final concern is how we should go about embedding human values into GenAI. Classically, this work has relied on formalizing human values into well specified “laws” (that strictly constrain the GenAI) or “reward functions” (that set goals for the GenAI), but both these classical approaches are thought to possess intractable problems due to the difficulty of avoiding misalignment between the specified goal or reward and our intended goal. . These approaches are thus susceptible to specification gaming, i.e., of behavior that achieves the goal or reward we specified without achieving our intended goal. If GenAI can develop more comprehensive cognitive capacities, then some have suggested that the study of human moral development could be usefully applied to GenAI.[6] On this view, GenAI would be aligned with human values through experiential learning in context-rich environments, similar to the ways that human learn Importantly, owing to our strengths in psychology and BCS, **UR has a comparative advantage in moral and developmental psychology and cognitive sciences.**

B. Work

The hype around GenAI is that it will fundamentally alter work across a large range of industries - including those that have historically relied on “white collar”, college-educated labor. By automating content creation (writing, image, music), data analysis (summarization, code generation), and other cognitive tasks, the promise is that these models can increase efficiency and output across sectors like media, arts, education, customer service, and management that have historically resisted widespread automation.

B1. Estimating GenAI's Economic Impact

A range of research has emerged over the past decade that attempts to predict the effect of AI on economic productivity and the availability and distribution of work.[7,8] However, the impact on the labor force is more complex. While generative AI may displace some routine jobs, it could also create new types of high-skilled roles focused on prompt engineering, model fine-tuning, and integrating AI systems into human workflows. The net effect on employment is still unclear, with predictions ranging from widespread job losses to a gradual shift towards human-AI collaboration. Likewise, the distribution of economic gains (nationally and globally) is unsettled. Methods to rigorously measure and predict these changes are sorely needed. Importantly, **UR has a small number of researchers in economics (Lisa Khan) and Simon (Rui Huaxui) who are already engaged in this work.**

B2. Training for the GenAI Economy

A range of research has emerged that attempts to identify the skills necessary for workers to either (i) successfully *collaborate* with GenAI in existing occupations, or (ii) *transition* to tasks that GenAI is unlikely to be able to automate. Some of this research overlaps with B1, and involves trying to predict which are the tasks where AI can help scaffold/supplement human work, and which are the tasks where AI will likely replace human work. A second strand asks what are the *educational* interventions that are necessary (at K-12, and post-college) to give students the skills the requisite skills. Importantly, owing to the Warner school and LIDA center, **UR has a comparative advantage in research into educational interventions around the use of technology.**

B3. Distributing the GenAI Dividend

A final question is how the (presumed) productivity gains should be distributed, especially in those contexts where AI is sophisticated enough that most economically-valuable tasks can be better performed by an AI than a human. In such a circumstance, the relative value of capital (GPUs, AI models, data centers) will rapidly swamp the value of labor, driving wages down. The research could draw from political philosophy, economic modeling of different UBI , assessment of UBI pilot studies (including one in the City of Rochester). To our knowledge, **UR currently does not have researchers focused on these questions.**

C. Trust

There is deep concern that GenAI's ability to quickly and cheaply generate realistic, compelling content will result in large changes to the way we come to know information, the ways in which we relate to other human beings, and the overall trust we have in political and cultural institutions.

Cross-cutting problems include: (1) identifying methods for measuring and monitoring social impact / trust, (2) how these issues intersect with existing social inequalities, (3) the *technical changes* that might build/protect trust (i.e. watermarking, transparency, etc), and (4) the *regulatory* and policy approaches to managing social trust.

C1. AI Bias and Representation:

Research focuses on how AI systems reflect and potentially amplify societal biases in language, images, and behavior. Studies examine representation in training data, how different groups are portrayed in AI-generated content, and the downstream effects on stereotype perpetuation. Key questions include how to measure bias in generative systems, methods for bias mitigation, and the broader societal impact of AI-perpetuated stereotypes.

C2. AI Deepfakes and Epistemological Crisis

This strand examines how synthetic media and AI-generated content challenge our collective ability to distinguish truth from fiction. At its most extreme, the idea is that a large amount of GenAI created content may lead people to not trust traditional forms of high quality evidence (i.e. video or audio of an event), Key questions include how *individual* deepfakes may be detected, how *widespread* use of GenAI may reshape information ecosystems, and what social solutions might help maintain shared reality (i.e. media literacy interventions).

C3. AI Companions and Human Relationships

This strand investigates how AI companions and chatbots influence human social development and relationship formation. Research examines whether AI relationships *substitute for* or *scaffold* human connections, how they affect emotional development (especially in children), and their impact on social skills and intimacy. Questions arise about attachment, authenticity, and the psychological effects of relationships with non-human entities. Importantly, **UR has a comparative advantage in this space because of our strong program in developmental psychology, social psychology, and child and adolescent mental health**.

D. Creativity

D1. Norms of Creativity with GenAI

One particularly important strand involves **analyzing evolving perceptions of creativity, originality, and artistic authenticity** given the advent of GenAI, including norms for authorship attribution, copyright frameworks, and intellectual property rights for AI-assisted works. This could include critically examining historical and emerging ethical and professional norms in music, literature, and the visual arts. These investigations could help develop guidelines for artists, platforms, and technologists, including consent mechanisms for training data, artist opt-out protocols, and labeling practices for AI generated content. This might require interdisciplinary workshops bridging legal, artistic, and technological perspectives to co-create governance models for generative AI in creative industries.

D2. Impact of GenAI on Creative Processes

A second strand would map the **effect of GenAI on creative workflow, ideation, and artistic production** across different media. This might include a variety of social scientific approaches, including:

- Comparative studies examining human-AI collaborative methodologies versus traditional solo creation.
- Investigate cognitive and psychological impacts of AI integration on artists and their creative processes.
- Quantitative and qualitative analysis of variations in creative outputs, including aesthetic diversity and evolution of genre.

But it might also involve critical humanistic methods that explore some of the assumptions and discourses around GenAI's effect on different creative disciplines (visual arts, music, writing) and uniquely integrate AI technologies. For instance, critical examination of claims about potential "democratization" and "accessibility", as well as investigations into "homogenization" of art, cultural biases and the "colonial standpoint" of GenAI systems. UR has a comparative advantage because of existing, externally-funded collaborative projects between the Eastman School of Music, the Warner School, Hajim and Arts and Sciences that focus on music technology and AI assisted learning. These projects include, the NSF funded TEAMuP project that explores the technological, educational and workforce related dimensions use of deep learning in music production.

D3. Teaching Creativity with GenAI

A third strand would investigate the **pedagogical approaches to integrating / resisting GenAI in creative arts education**. This might involve designing and evaluating curriculum models for exploring AI as a collaborative tool or creative catalyst, but it might also involve identifying when to limit or forbid GenAI use in order to grow foundational creative skills. This dovetails with research on alternative assessment frameworks for creative conduct that recognizes the value/existence of GenAI while meeting pedagogical goals. Some of this might involve studying the psychological and motivational impacts of GenAI on artistic learning environments. UR has a comparative advantage because of existing strengths in Music Education that incorporates AI technology, (Sangmi Kang, ESM), AI for Education (Zhen Bai), as well as strong faculty interest in Writing, Speaking, Argument.