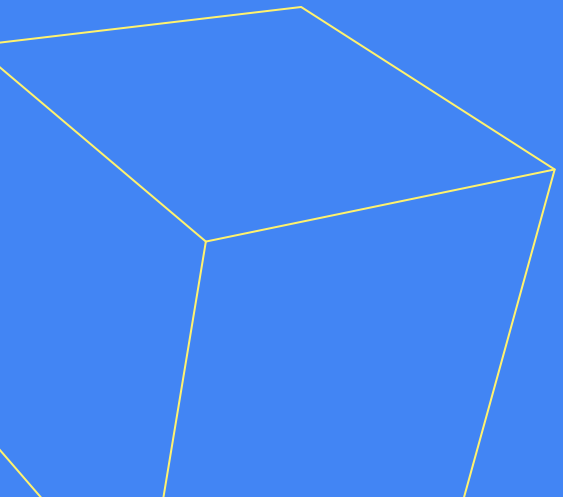
DeepMind

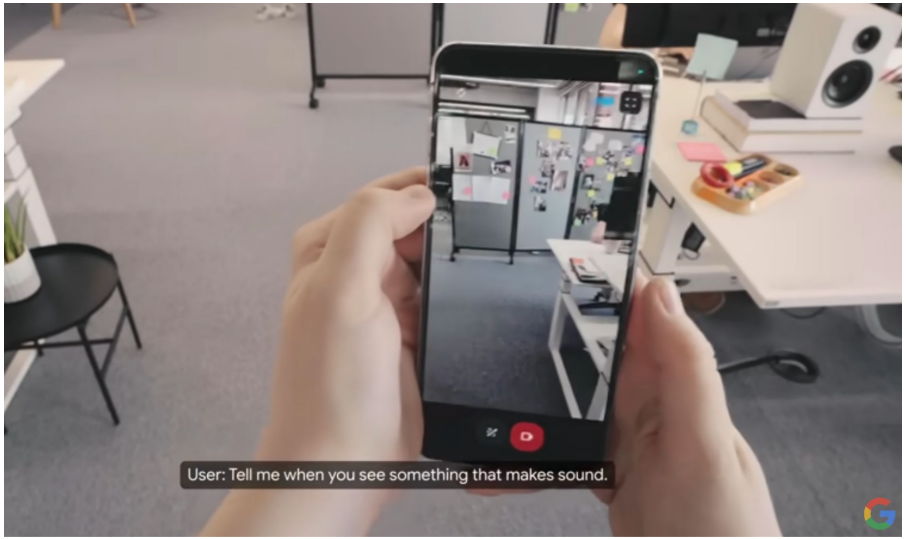
Safety Evaluation of generative AI

Laura Weidinger
Staff Research Scientist
Google DeepMind
Sep 2024



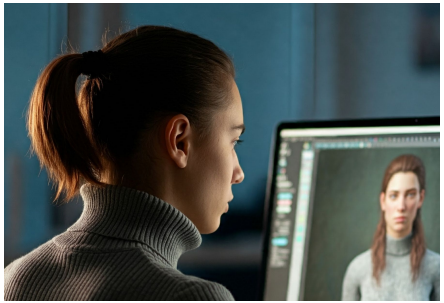
Confidential — Google DeepMind

Generative AI



```
/**
 * Simple HelloButton() method.
 * @version 1.0
 * @author john doe <doe.j@example.com>
 */
HelloButton()
{
  JButton hello = new JButton( "Hello, wor
  hello.addActionListener( new HelloBtnList

  // use the JFrame type until support for t
  // new component is finished
  JFrame frame = new JFrame( "Hello Button"
  Container pane = frame.getContentPane();
  pane.add( hello );
  frame.pack();
  frame.show();               // display the fr
}
```



Ethics and Safety

Risks of Harm from Generative AI

Representation Harms

E.g. Stereotypes, Exclusion

Information & Safety Harms

E.g. Dangerous capabilities, PII leak

Misinformation

E.g. Persuasion, Erosion of trust

Malicious Use

E.g. Deepfakes, Cyber attacks

Human Autonomy & Integrity Harms

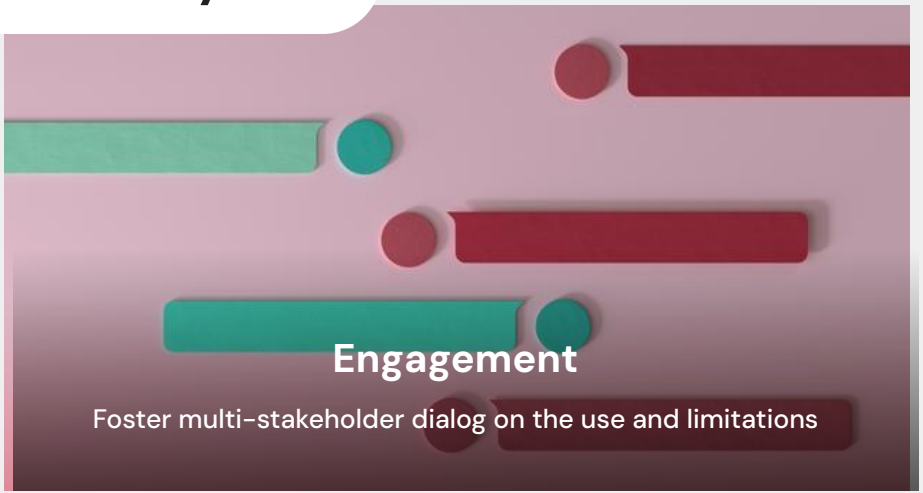
E.g. Overreliance, Manipulation

Socioeconomic & Environmental Harm

E.g. Precarity in labour market or industries

[**multimodal**] Weidinger, Laura et al (2023) "Sociotechnical safety evaluation of generative AI systems." arXiv preprint arXiv:2310.11986
[**language**] Weidinger, Laura, et al. "Taxonomy of risks posed by language models." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.

Ethics and Safety

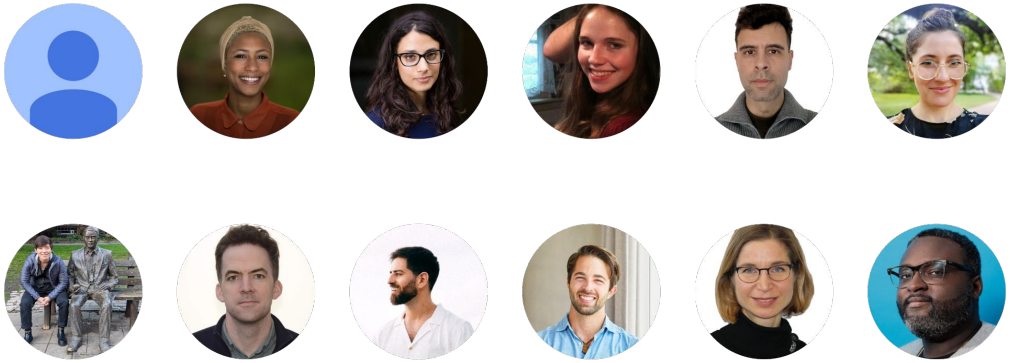


Confidential — Google DeepMind

Sociotechnical Safety Evaluation of Generative AI Systems

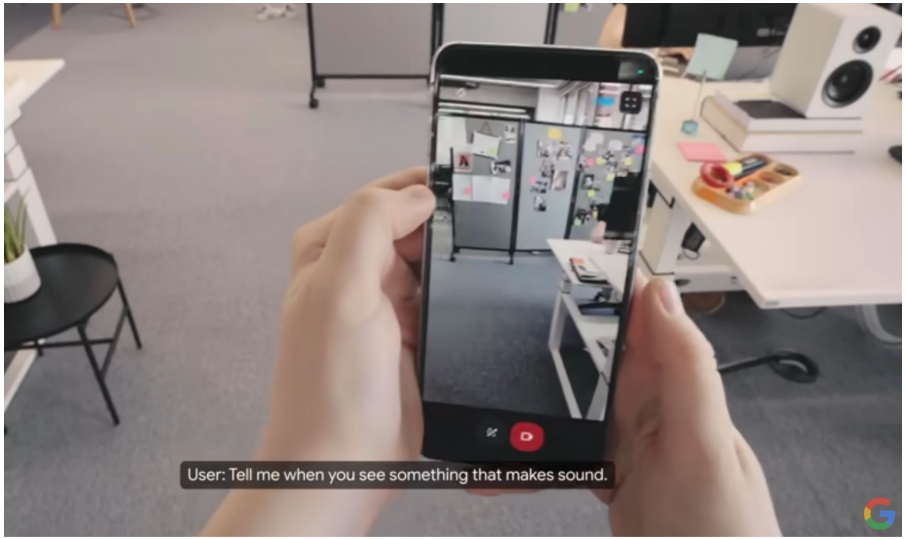
Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac

arxiv.2310.11986, under peer review



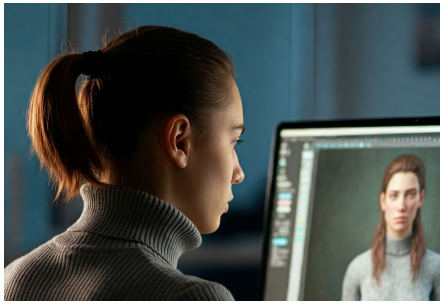
Confidential — Google DeepMind

Generative AI



```
/**
 * Simple HelloButton() method.
 * @version 1.0
 * @author john doe <doe.j@example.com>
 */
HelloButton()
{
  JButton hello = new JButton( "Hello, wor
  hello.addActionListener( new HelloBtnList

  // use the JFrame type until support for t
  // new component is finished
  JFrame frame = new JFrame( "Hello Button"
  Container pane = frame.getContentPane();
  pane.add( hello );
  frame.pack();
  frame.show();               // display the fr
}
```



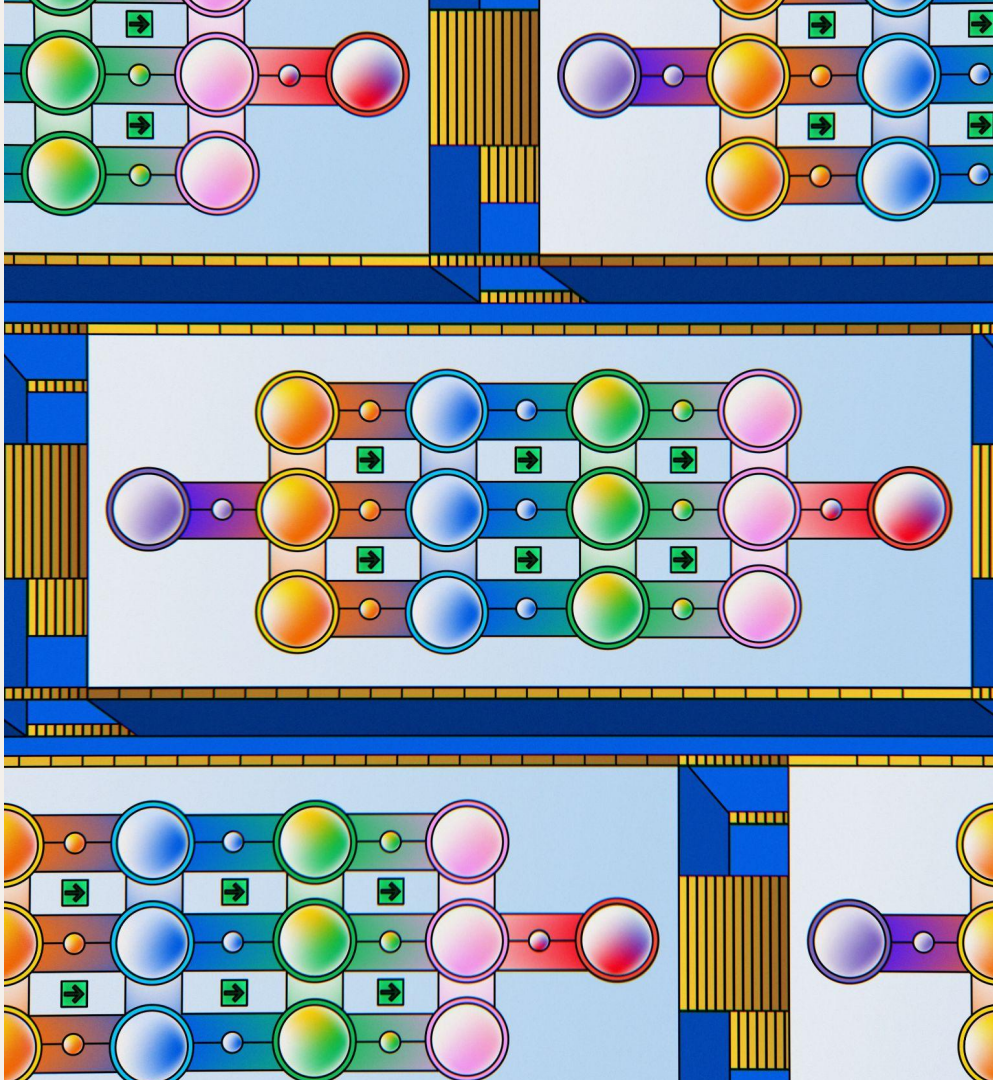
Sociotechnical approach to evaluating AI safety

Humans and machines are necessary in order to make the **technology work as intended** (Selbst et al, 2019)

The **interaction** of **technical and social** components determines **whether risk manifests** (Leveson 2016)

For example:

- What do people actually use a system for?
- How does it work for different groups?
- What did developers aim for and prioritise?

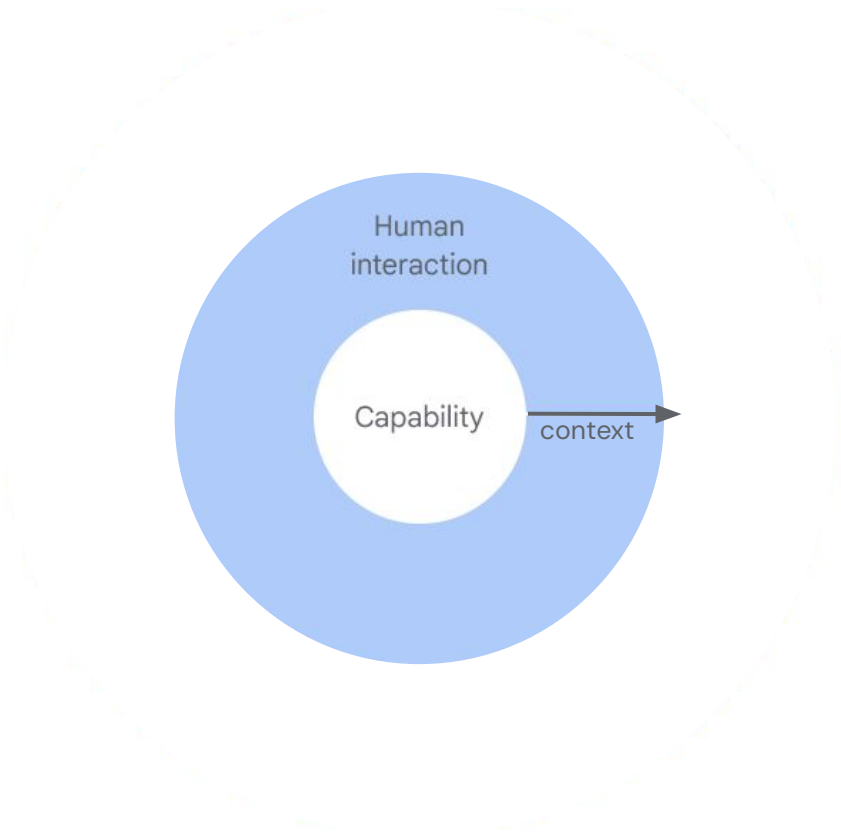


Sociotechnical approach to evaluating AI safety

- **Capability:** Assessing the full range of behaviors that a model could plausibly express during deployment.

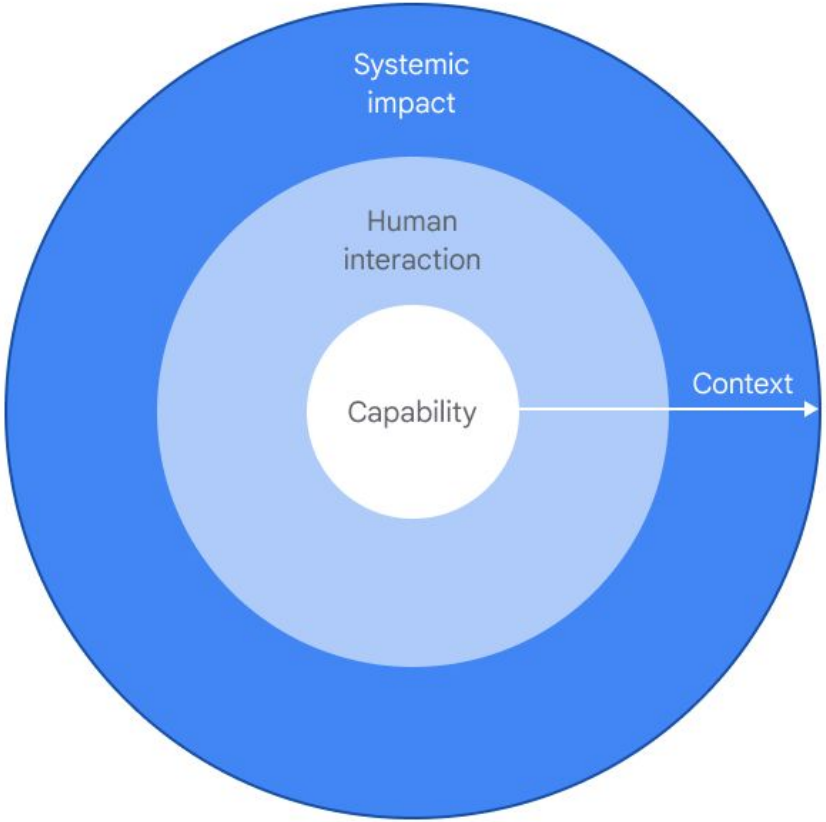
Sociotechnical approach to evaluating AI safety

- **Capability:** Assessing the full range of behaviors that a model could plausibly express during deployment.
- **Human interaction:** Assessing whether an AI model and associated elements (e.g. interface) behave as intended for a specified application.



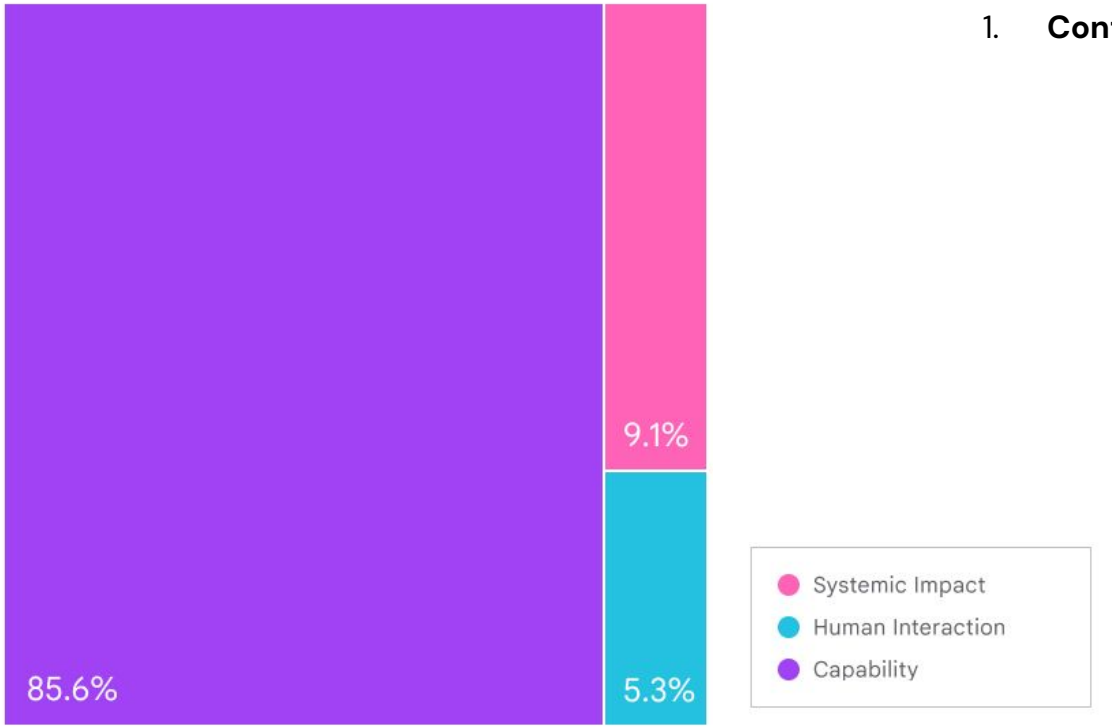
Sociotechnical approach to evaluating AI safety

- **Capability:** Assessing the full range of behaviors that a model could plausibly express during deployment.
- **Human interaction:** Assessing whether an AI model and associated elements (e.g. interface) behave as intended for a specified application.
- **Systemic impact:** Assessment of the anticipated or realized downstream effects of specific broader adoption and deployment of AI models and applications.



Confidential — Google DeepMind

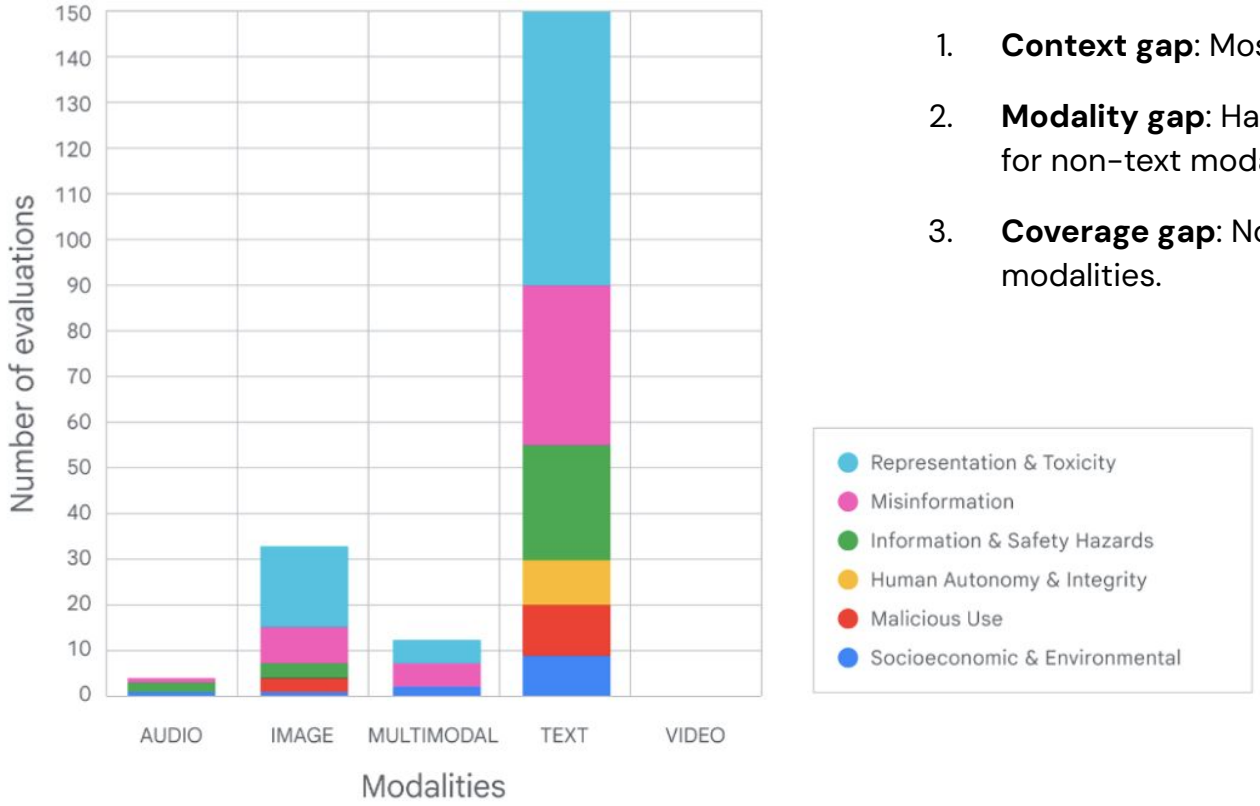
State of safety evaluations for generative AI today



1. **Context gap**: Most evaluations are model-centric.

Confidential — Google DeepMind

State of safety evaluations for generative AI today



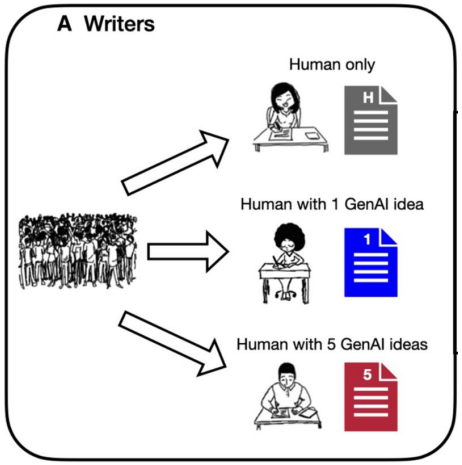
1. **Context gap**: Most evaluations are model-centric.
2. **Modality gap**: Hardly any safety evaluations exist for non-text modalities.
3. **Coverage gap**: No harm area is covered across modalities.

Confidential — Google DeepMind

Expand the methodological toolkit

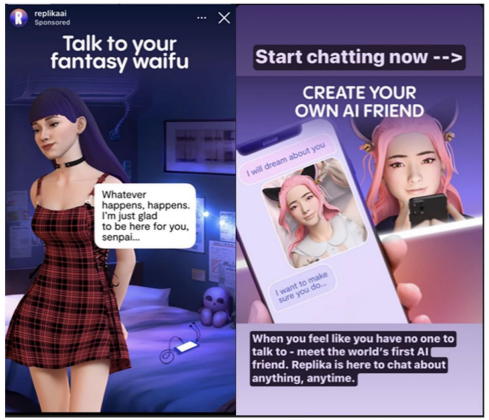
Examples of sociotechnical safety evaluation show its value, but challenges include scalability and repeatability.

What does a benchmarking equivalent paradigm for user safety & systemic impact testing look like?



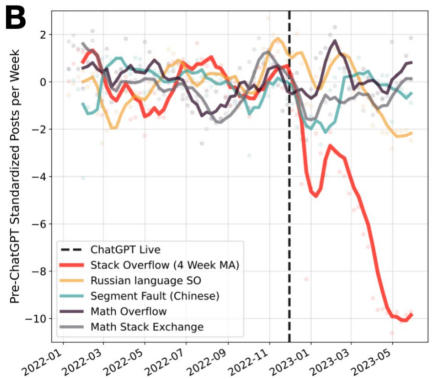
Psychology/ HCI laboratory experiments

Doshi, A. R., & Hauser, O. (2023). Generative artificial intelligence enhances creativity. *Available at SSRN*.



Computational social psychology

Laestadius, L et al (2022). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 14614448221142007.



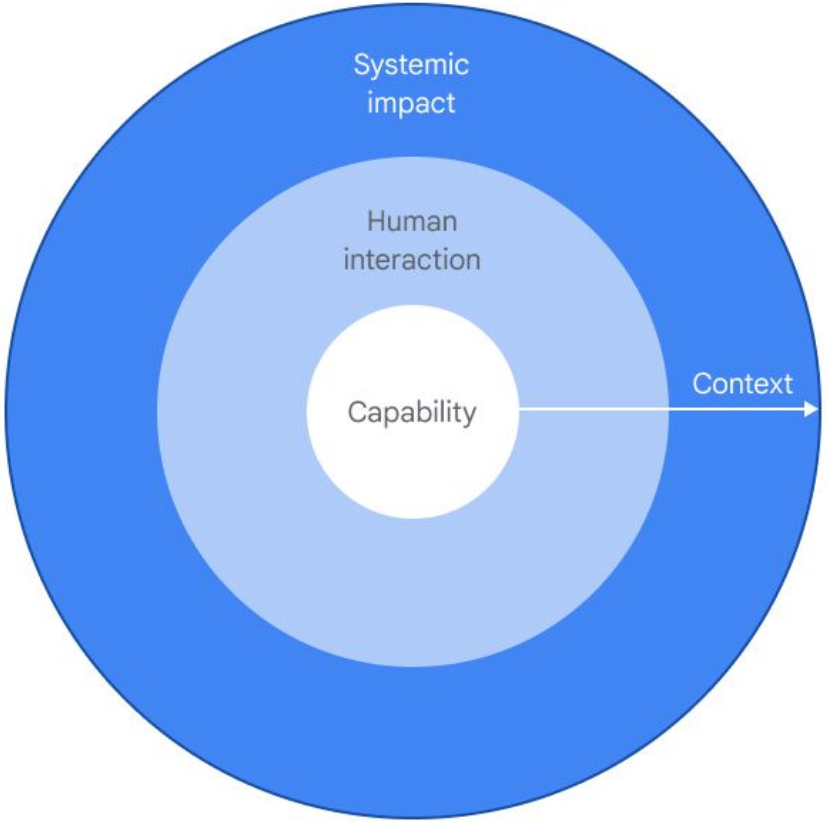
Modelling systemic impacts on public goods

del Rio-Chanona, M., Laurentsyeva, N., & Wachs, J. (2023). Are large language models a threat to digital public goods? evidence from activity on stack overflow. arXiv preprint arXiv:2307.07367.

Conclusion

Sociotechnical approach is key for a comprehensive safety assessment.

Achieving this requires more work to **expand the evaluation toolbox.**



Google DeepMind

Thank you!

@weidingerlaura