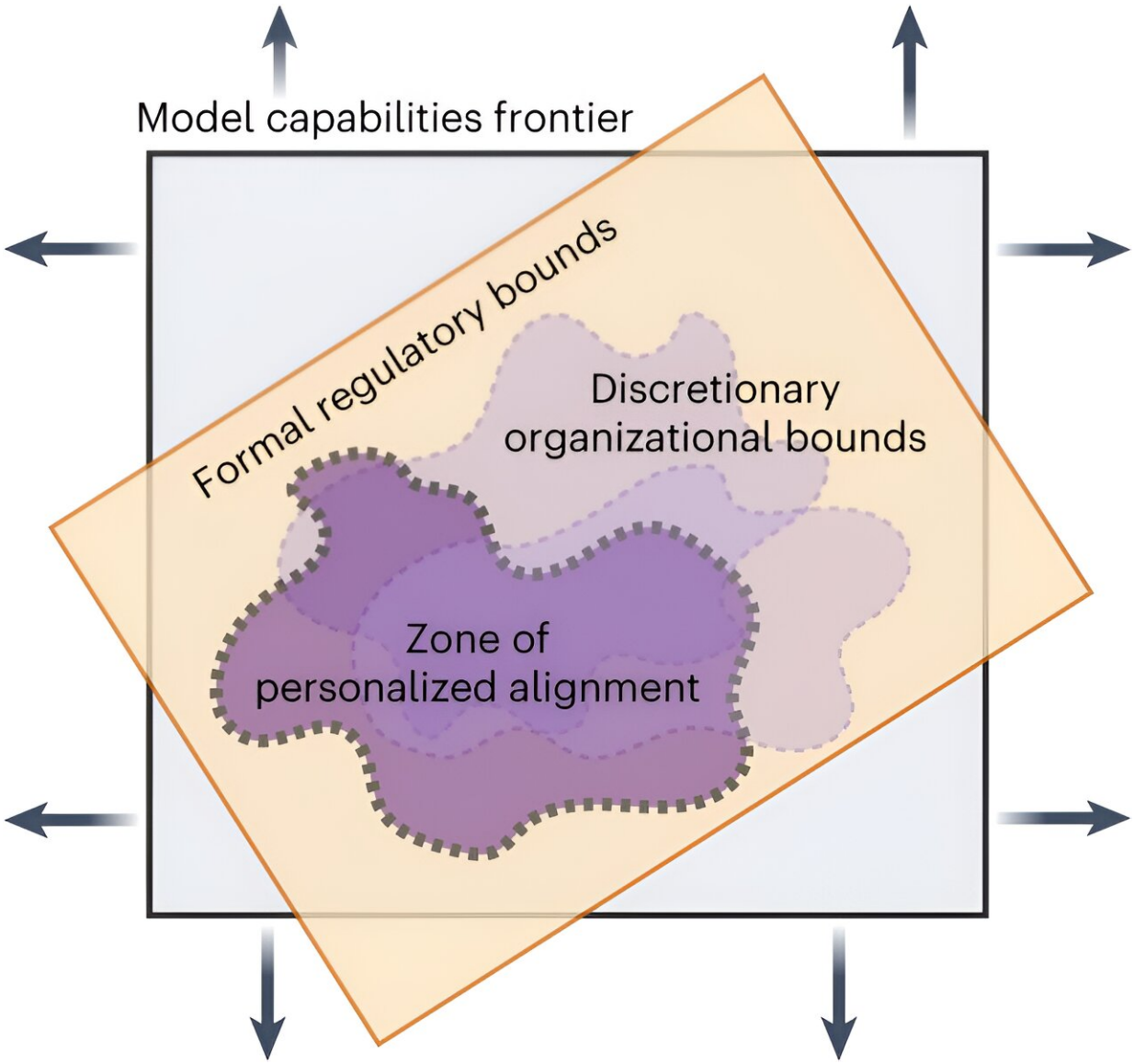
# Personalization has the potential to democratize who decides how LLMs behave

April 23 2024




Hierarchical bounds on personalized alignment. Credit: *Nature Machine*

*Intelligence* (2024). DOI: 10.1038/s42256-024-00820-y

A new paper from researchers at Oxford Internet Institute, University of Oxford, highlights the benefits and risks of personalizing Large Language Models (LLMS) to their users.

LLMs are artificial intelligence systems that generate written responses to text prompts. Due to their stratospheric growth in the past two years, hundreds of millions of people now interact with LLMs. Yet, the initial design and development decisions behind LLMs mean that small groups of developers, researchers or human annotators are providing the technology with the information it needs to respond to queries.

This influences the conversational norms, values or political beliefs embedded in a model. Invariably, without wider participation during training, there is a risk that the diverse worldviews of those who use LLMs are excluded or misrepresented in their text responses.

In their paper, published today (April 23) in *Nature Machine Intelligence*, the Oxford researchers present personalization as a potential solution to sustaining different worldviews in language technologies.

Like any new technology, they argue the responsible adoption of personalization requires balancing the new benefits it can bring while managing potential risks for individual users and society as a whole. These benefits and risks are not purely theoretical: it has recently become possible to personalize ChatGPT, a widely-known LLM developed by OpenAI.

For individuals, the benefits of personalization include increased ease in finding information, in a format tailored to their communication

preferences. The user may also enjoy a technology that better adapts to their diverse beliefs or memorizes information about their needs. Personalization may result in a more empathetic connection and a sense of ownership of it being "my technology."

However, this greater usefulness and deeper connection to the technology may fuel over-dependence and addiction. As with other types of artificial intelligence (AI), there is the risk of people anthropomorphizing the technology and becoming attached to it. Personalization is not possible without personal data; so, there is also an increased risk of users' privacy being compromised.

Personalizing LLMs also impacts society. Personalization can bring better inclusivity and democratization by diversifying which members of society have influence on how LLMs behave. If answers are more in tune to individuals' needs, labor forces using LLMs could also become more productive.

However, not everyone has equal access to technology and those excluded risk becoming more disadvantaged by a widening digital divide. A further concern is that personalization could contribute to societal polarization and echo chambers when individuals less frequently encounter beliefs different from their own. The technology also has the potential to become a powerful instrument in generating persuasive and targeted disinformation, which is already problematic in the online world.

Commenting on the findings Hannah Rose Kirk, lead author and DPhil Student at Oxford Internet Institute, University of Oxford, said, "It's vital we start the conversation now on what responsible personalization looks like, as the technology is being developed. That way we have the best chance of enabling individuals and society to reap its benefits, without a lag in understanding or regulating the risks."

Professor Scott A. Hale, Oxford Internet Institute, University of Oxford added, "Examining the risks and benefits of personalization now while approaches are still being developed is the best way to create more inclusive and responsible technologies."

Dr. Bertie Vidgen, Visiting researcher at the OII and The Alan Turing Institute, and co-supervisor and author, added, "Personalized AI models feel like an obvious win—but that's true only up to a point. If we aren't attuned to the risks as well as the benefits, the consequences could be huge. This paper brings some much-needed clarity to this important debate."

**More information:** Hannah Rose Kirk et al, The benefits, risks and bounds of personalizing the alignment of large language models to individuals, *Nature Machine Intelligence* (2024). DOI: 10.1038/s42256-024-00820-y

Provided by University of Oxford