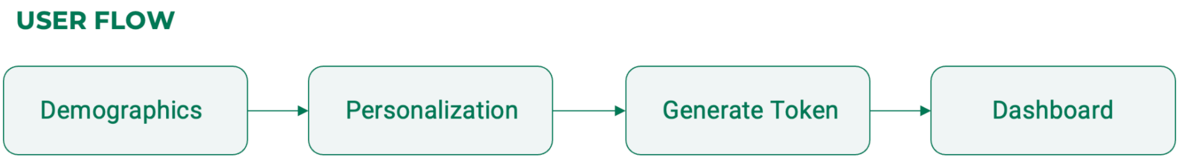
EthicsNet

Personalized Fine-Tuning Token for AI Value Alignment

Really Simple Basic AI Value Alignment for Everyone

Summary

We're working on a new system that makes it easier for artificial intelligence to understand what's important to you personally, while also reducing unfair or biased decisions. Our system includes easy-to-use tools that help you identify and mark different situations where the AI might be used. These tools use special techniques, like breaking down text into meaningful parts and automatically labeling them, to make it simpler to create settings that are tailored to you. By doing this, we aim to address the problem of AI not fully grasping people's unique backgrounds, preferences, and cultural differences, which can sometimes lead to biased or unsafe outcomes.


USER FLOW

Demographics → Personalization → Generate Token → Dashboard

The non-summary

We've done an extensive review of existing research on automated methods for adding notes or labels to data, as well as developing a frameworks for tailoring AI to individual values. Based on this, we're suggesting a practical implementation.

Our approach lets us label different human behaviors in various settings, like social media, in a way that reflects societal values. For example, aggressive or bullying behavior can be marked as "not acceptable," while positive actions can be labeled as "good," showing what society generally approves or disapproves of.

These labels don't just identify behaviors; they also connect them to the cultural norms from which they come. This creates a sort of moral map that reflects the views of different groups of people. Importantly, those adding the labels will indicate their own value preferences, rather than telling others what they should think or do.

The beauty of this approach is its flexibility: it can act as a safety net against the dangers of getting stuck in a single set of values that may not be universally applicable. When this labeled data is used to train a language model, the AI can better align with societal values. We can even include information on the outcomes of certain behaviors, guiding the AI to make value-based decisions that are generally considered good or bad by society.

This method can also improve existing techniques like Reinforcement Learning from Human Feedback (RLHF), which is a promising but difficult approach to train AI. The company Hugging Face points out that collecting quality human feedback for RLHF is both time-consuming and expensive. Our approach could streamline this by providing high-quality, nuanced labels.

Looking ahead, the data we collect could serve as a basis for creating scenarios that help users quickly adjust an AI's behavior according to a simplified set of value choices, like "do you prefer option A, B, or C?" This could be a game-changer for making AI safer and better aligned with human values. These scenarios could even act as "test cases" to check an AI system's moral fitness in specific situations.

> *"…You, as a user, should be able to write up a few pages of 'here's what I want; here are my values; here's how I want the AI to behave' and it reads it and thinks about it and acts exactly how you want because it should be your AI."*
>
> *– Sam Altman, OpenAI*

De-risking Value Misalignment

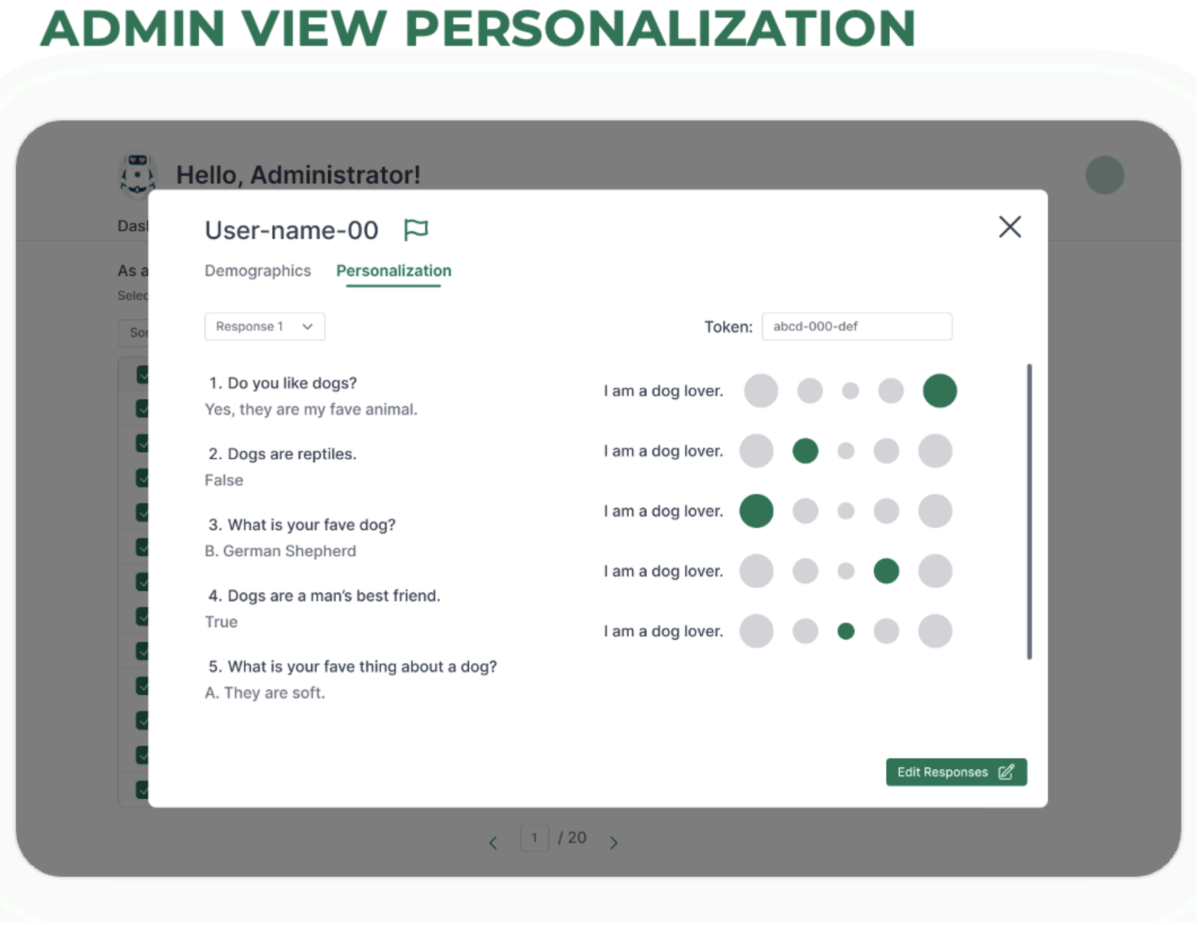
Current AI models run into problems because they often lack good-quality data that reflects the wide range of human values across different cultures and situations. Simply adding more data may make the AI more capable, but not necessarily more reliable or trustworthy. This can be a major issue, especially when the AI is making important decisions that could affect people's lives. Using a one-size-fits-all set of "safe" values to guide the AI can also be problematic, as this could unfairly favor certain groups over others, such as those in the Global North over those in diverse local settings.

Ignoring the complexity of human values can lead to significant risks. For example, an AI designed to strictly follow a certain ideology might overlook critical information, potentially leading to dangerous outcomes.

A better approach could involve collecting a wide variety of human values and identifying common ground that most people can agree on most of the time. This could make AI not only more empathetic but also more effective in helping humans coordinate their actions.

Importantly, we can embed this collective understanding of human values into the AI to improve what's called "alignment of intent"—making sure the AI's underlying goals match those set by humans. This is different from simply training the AI to perform specific tasks well.

Adding information about the consequences of different actions can further refine the AI's understanding. This helps it not only to distinguish right from wrong but also to understand why something is considered right or wrong. If the training data effectively captures this nuanced range of human values, the AI is more likely to apply these values appropriately in new situations it wasn't specifically trained for, essentially learning a model of morality and consequences. This is crucial for aligning the AI's intent with human-defined goals.



ADMIN VIEW PERSONALIZATION

Hello, Administrator!

User-name-00

Token: abcd-000-def

1. Do you like dogs?
Yes, they are my fave animal.

2. Dogs are reptiles.
False

3. What is your fave dog?
B. German Shepherd

4. Dogs are a man's best friend.
True

5. What is your fave thing about a dog?
A. They are soft.

I am a dog lover. ○ ○ ○ ○ ●

I am a dog lover. ○ ● ○ ○ ○

I am a dog lover. ● ○ ○ ○ ○

I am a dog lover. ○ ○ ○ ● ○

I am a dog lover. ○ ○ ● ○ ○

The Task Scope

After successfully creating a blueprint for tailoring AI to individual values in our last project phase, we're now focused on actually building out further. We've already started developing the user interface and the underlying technology to make this possible. However, there are still some components that need work. For example, we plan to quantify values on a 1-5 scale using a Likert Scale, and we need to figure out how to apply these numerical values to adjust AI behavior effectively.

Given that we're dealing with sensitive information about people's value preferences, strong cybersecurity measures are also a priority. While we've put some basic security practices in place, we aim to make our system as secure as possible.

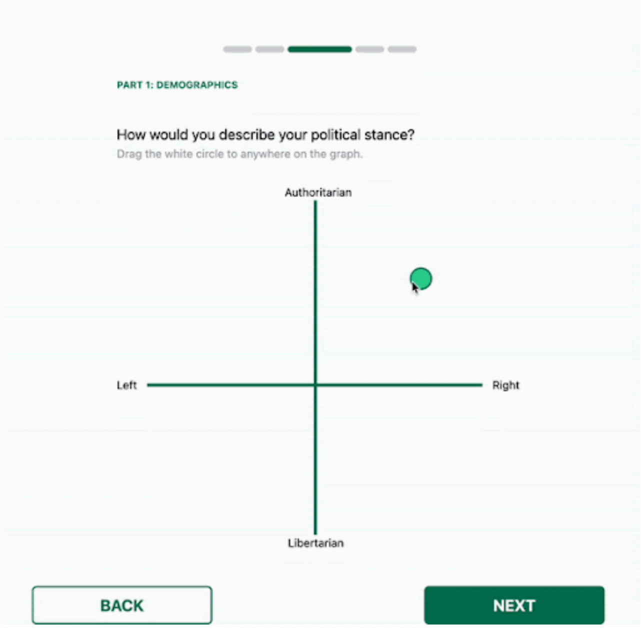
We're looking for experts in Vectorization, Fine-Tuning/RLHF, and Cybersecurity to join us. Our team will have weekly meetings to ensure quick progress, with the goal of developing systems that are ready for real-world testing. We aim to offer people a more personalized and context-aware interaction with AI.

Output

By the end of our project phase, we plan to launch a new system focused on making AI understand individual behaviors and values better. The first thing we'll release is a *web-based tool that allows people to customize AI according to their values*. This will be followed by *a set of tools that anyone can use to add notes or labels to behaviors* (annotation), thereby helping AI understand such behaviors better. These tools will be user-friendly, thanks in part to chat-based interfaces and features like automatic labeling.

Our main goal is to create a digital token that captures your personal values, which can then be shared with other AI systems via an online interface (API). A secondary goal is to demonstrate how our tools make the complex task of adding these notes or labels much simpler, enabling more people to participate in shaping how AI understands human values.

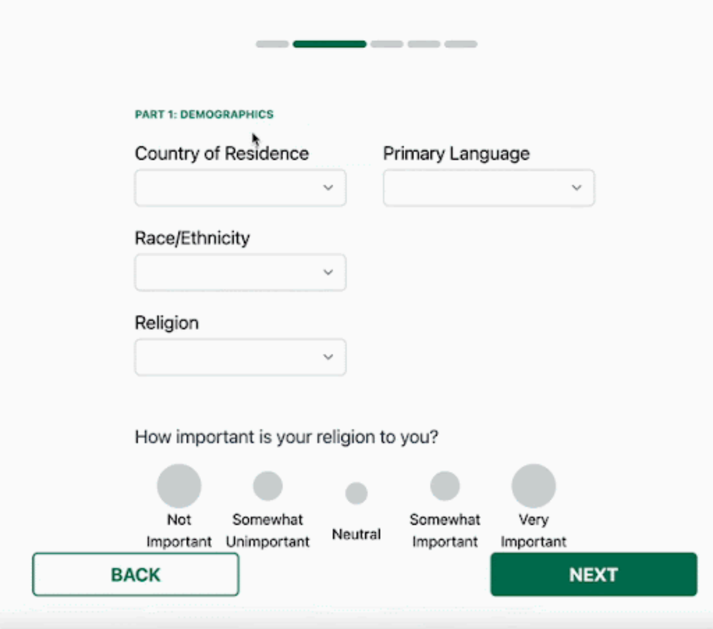
We'll document all of these developments in a research paper, aiming for publication in a specialized AI journal. This will provide a detailed account of what we've achieved and how we did it, contributing to ongoing efforts to align AI with global values.



PART 1: DEMOGRAPHICS

How would you describe your political stance?

Drag the white circle to anywhere on the graph.



PART 1: DEMOGRAPHICS

How important is your religion to you?

Risks and downsides

While the main focus of our project is to improve how AI understands and aligns with individual behaviors and values, we're fully aware of the cybersecurity risks involved. We've put basic security measures in place and are actively seeking experts to make our system even more secure.

When it comes to collecting data, privacy is a top concern. All participants will have the option to remain anonymous or use a pseudonym, and any personal information will be kept separate from the data on their values. This will make it harder for anyone to link the data back to specific individuals. We'll also consider breaking the data into smaller pieces, known as "sharding," based on different contexts or cultures. We plan to strictly follow GDPR and other privacy laws, and we'll make sure our data storage systems are highly secure.

To make sure our system understands a wide range of human values, we're going to include participants from diverse backgrounds, covering different demographics, political views, and geographic locations. The system will also be accessible, supporting multiple languages and devices.

To avoid the risk of people assuming their opinions are more widely held than they actually are—a phenomenon known as the False Consensus Effect—we'll use pseudonyms. To filter out unreliable or biased inputs, we'll use a peer-review and karma system, inspired by platforms like Wikipedia that successfully manage collaborative efforts.

The broader research project, of which this is a part, has received ethical approval from my University and is part of my Doctoral Research.

Acknowledgements

Sincerest thanks to Dr. Linda Linsefors, Dr. Koen Holtman, Remmelt Ellen, Paul Bricman, Chris Leong, and an anonymous editor for highly insightful feedback in this proposal. Thanks to Lukas Petersson and Benjamin Sturgeon who made very valuable contributions to this project by developing new lines of research, shaping the agenda, and specifying key challenges to be tackled in future work.

Team

Team size
The team of three people currently is based in the United Kingdom. The project is headquartered in the UK, with supervisory support from The University of Gloucestershire, and support also offered from the University of Illinois's Hack 4 Impact program. We are looking for 2-3 further team members to assist this with the implementation of the personalization and annotation framework.

Research Lead
Eleanor 'Nell' Watson nell@nellwatson.com Based in Northern Ireland (my Resume & LinkedIn). I anticipate devoting 20-30 hours per week to this project.

I have background working knowledge of machine vision and several aspects of AI ethics and safety. For the past two years I have been deeply steeped in research on annotation of behavior, and have a Systematic Literature Review published on this subject, as well as a forthcoming paper on the framework itself accepted for publication in a high-impact journal. I have also been considering designs for a workflow for a new generation of behavioral annotation tools which will be extremely easy to use by the general public, to a level not feasible until very recently.

Skill requirements:

Team members should be (at least somewhat) knowledgeable in at least one of the below areas:

1. **Fine-Tuning/RLHF/Constitutional AI Expertise**
   We're interested in how our "values token" can be effectively used with techniques like Reinforcement Learning from Human Feedback and Fine Tuning. Our eventual goal is to produce a token output that can be directly interfaced with third-party models, though significant further research is required to interface value preferences (textual or vector) with machine learning models to substantially alter their behavior, especially in a safe and robust manner. Theoretical approaches in representation engineering and contrastive preference modeling highlight potential means to accomplish this, and perhaps part-trained model approaches also, but significant uncertainty remains.

2. **Cybersecurity Expertise**
   Given that we'll be collecting potentially sensitive data about people's values, cybersecurity is a top concern. We're looking for experts who can help make our system as secure as possible, including methods such as homomorphic encryption.

3. **Vector Databases**
   We plan to turn Likert-scale responses about values into numerical vectors (think of it as 'value2vec'). We need people who are skilled in creating these kinds of vector databases.

4. **Product Management**
   What would product/market fit look like for personalized AI value alignment? How to make it more compelling? What tests could improve knowledge? What do LLM mavens want/need? What's the MVP for a researcher preview? How best to build feedback loops, integrating those into the product?

If you have expertise in any of these areas, we'd love to have you on the team to help us advance this important work. Please contact nell@nellwatson.com for further information. Thank you.