CS224V: Ethics and Policy Lecture

Value-Based LLMs

How Bias, Alignment, and your values come together

Tommy Bruzzese

Wed Nov 29th, 2023

CS224V: Ethics and Policy Lecture

Value-Based LLMs

How Bias, Alignment, and your values come together

Tommy Bruzzese

Wed Nov 29th, 2023

AI Ethics started with fairness & bias

AI Ethics started with fairness & bias

Are Emily and Greg More Employable than Lakisha and Jamal? [Bertrand & Mullainathan '03]

- White names receive 50% more callbacks for interviews

AI Ethics started with fairness & bias

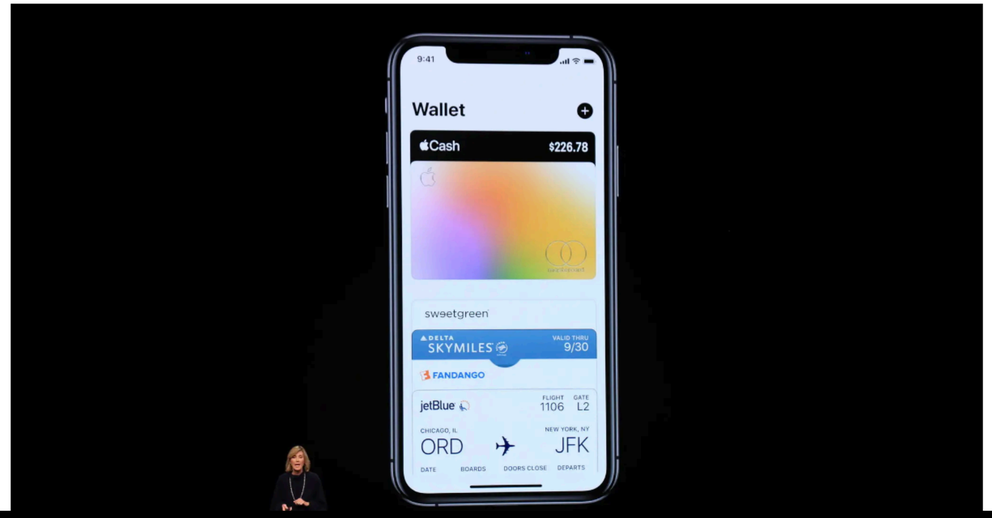
- COMPAS is **2x as likely** to *mislabel* Black individuals who did not go on to recommit crime incorrectly as risky than white individuals
- COMPAS *mislabeled* white individuals who did go on to recommit incorrectly as recommended-for-release more so than Black individuals

AI Ethics started with fairness & bias

Apple Card Investigated After Gender Discrimination Complaints

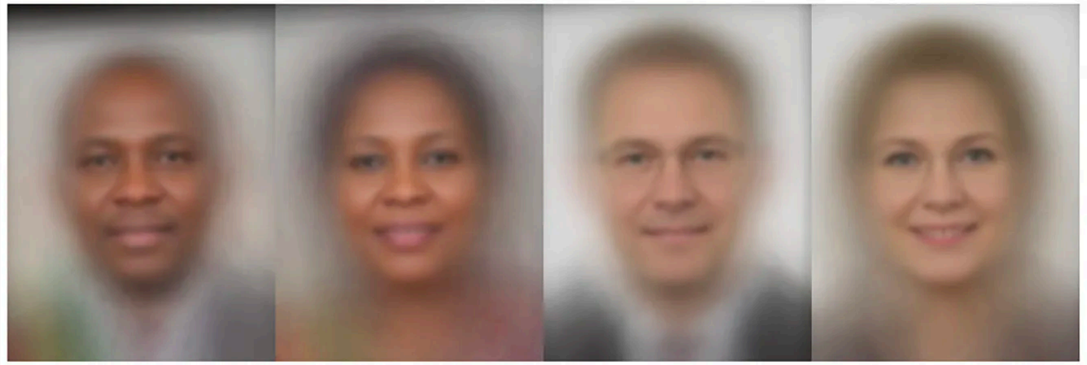
A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.

Share full article



AI Ethics started with fairness & bias

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



[Buolamwini and Gebru, "Gender shades," 2018]

AI Ethics started with fairness & bias

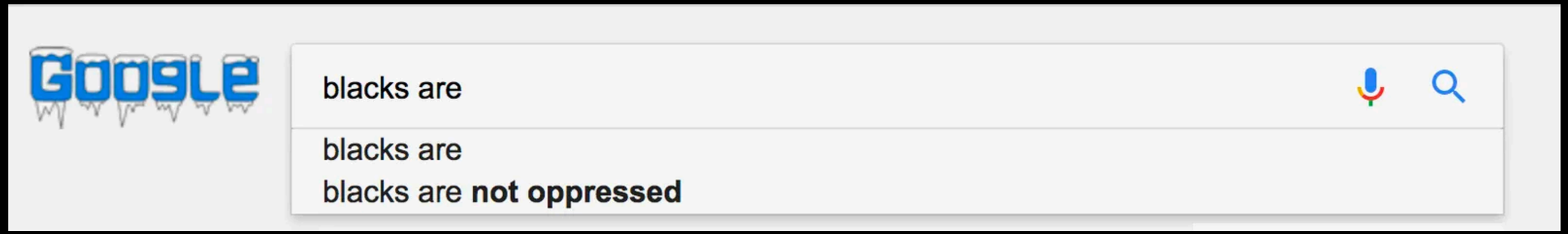
Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

2016

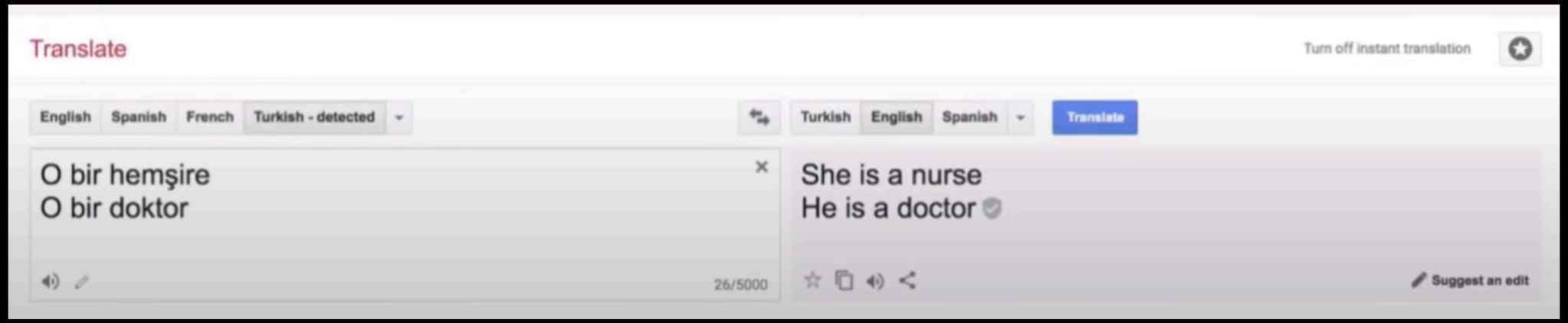
man:doctor :: woman:nurse

AI Ethics started with fairness & bias



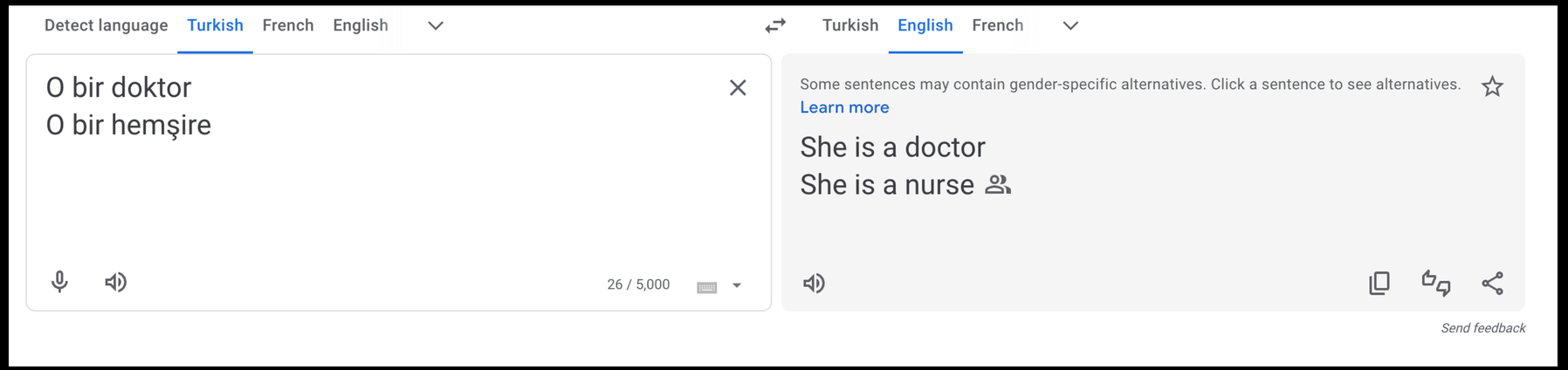
[*Wired*, “Google Autocomplete Still Makes Vile Suggestions” 2018]

AI Ethics started with fairness & bias



[Crawford, “The Trouble with Bias” 2017]

AI Ethics started with fairness & bias



[Author, 2023]

Harms in decision-making and in representation

[Crawford, “The Trouble with Bias,” 2017]

Allocational

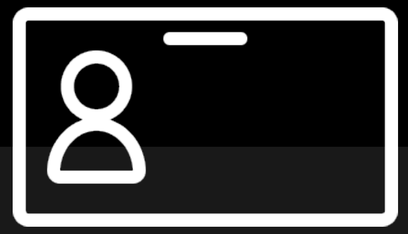
Resources, jobs

Harms in decision-making and in representation

[Crawford, “The Trouble with Bias,” 2017]

Allocational

Resources, jobs



Representational

Identity, stereotype

THE PROBLEM WITH BIAS

Allocation	Representation
Immediate	Long term

[Crawford, “The Trouble with Bias,” 2017]

THE PROBLEM WITH BIAS

Allocation	Representation
Immediate	Long term
Easily quantifiable	Difficult to formalize

[Crawford, “The Trouble with Bias,” 2017]

THE PROBLEM WITH BIAS

Allocation	Representation
Immediate	Long term
Easily quantifiable	Difficult to formalize
Discrete	Diffuse

[Crawford, “The Trouble with Bias,” 2017]

THE PROBLEM WITH BIAS

Allocation	Representation
Immediate	Long term
Easily quantifiable	Difficult to formalize
Discrete	Diffuse
Transactional	Cultural

[Crawford, “The Trouble with Bias,” 2017]

“Many of these failures have been the direct result of ***underrepresentation***, ***misrepresentation***, or the complete ***lack of representation*** of these groups ***in the data*** upon which these systems are built (Paullada et al., 2020)”

[Denton, 2021]

“Garbage in, garbage out”

Datasheets for Datasets

Separation

Independence

Model cards

Counterbalancing

Adversarial examples

Data-scrubbing

Sufficiency

Datasheets for Datasets

Separation

Independence

Model cards

Counterbalancing

Adversarial examples

Data-scrubbing

Sufficiency

“Relatively little work has been done to examine the ***norms***, ***values***, and ***assumptions*** embedded”

[Denton, 2021]

RESEARCH-ARTICLE OPEN ACCESS

The Values Encoded in Machine Learning Research

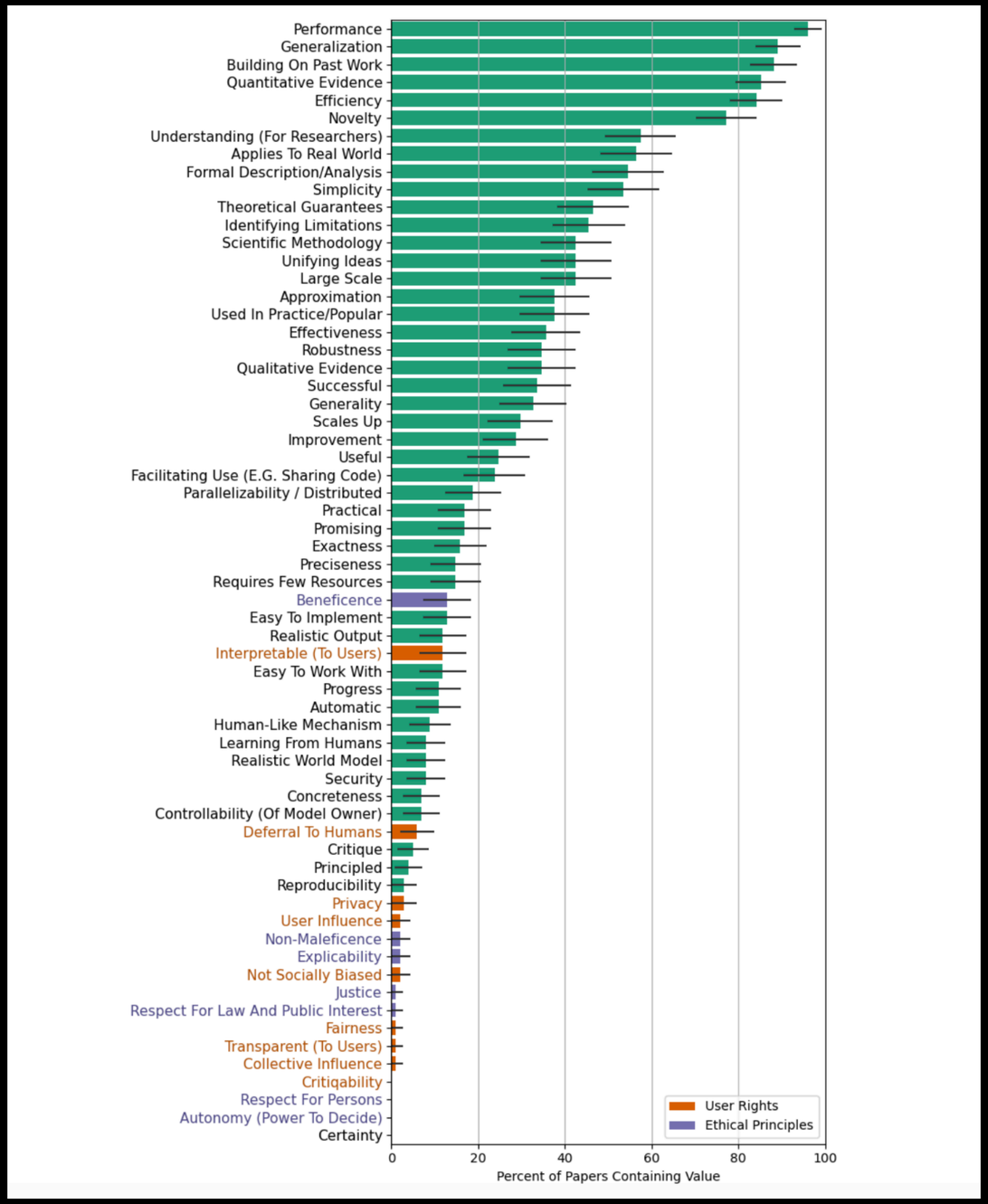
Authors: Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, Michelle Bao

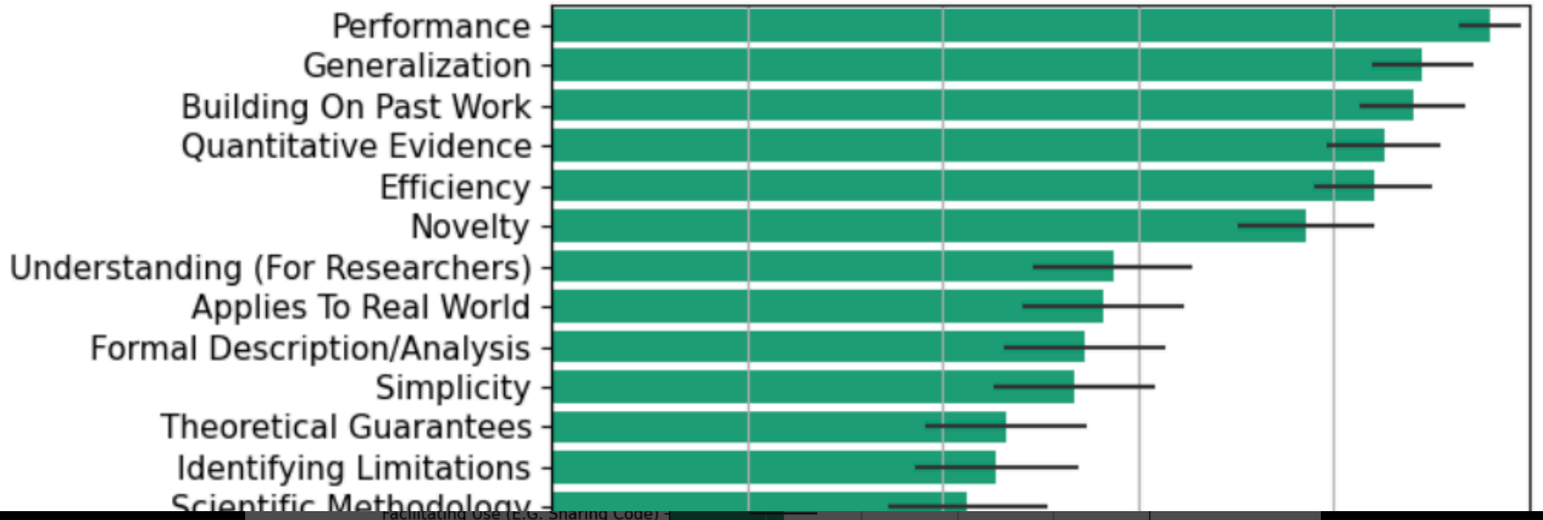
Authors Info & Claims

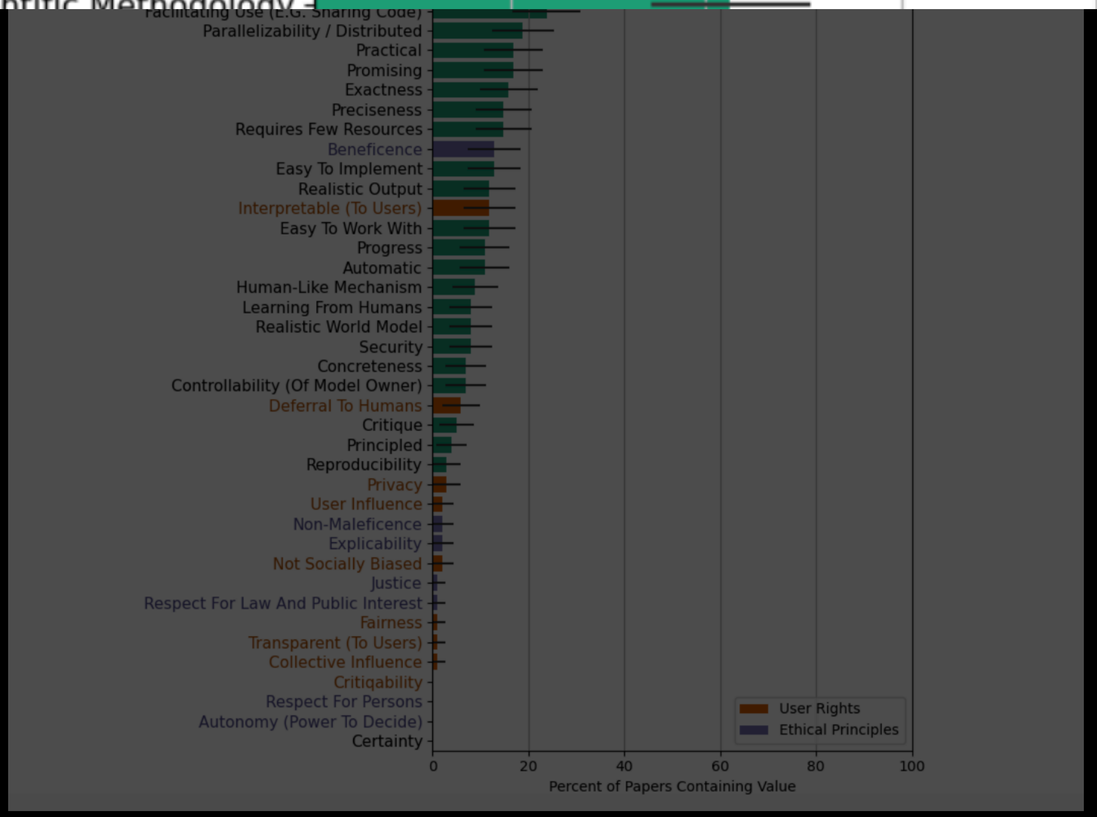
FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency • June 2022 • Pages 173–184

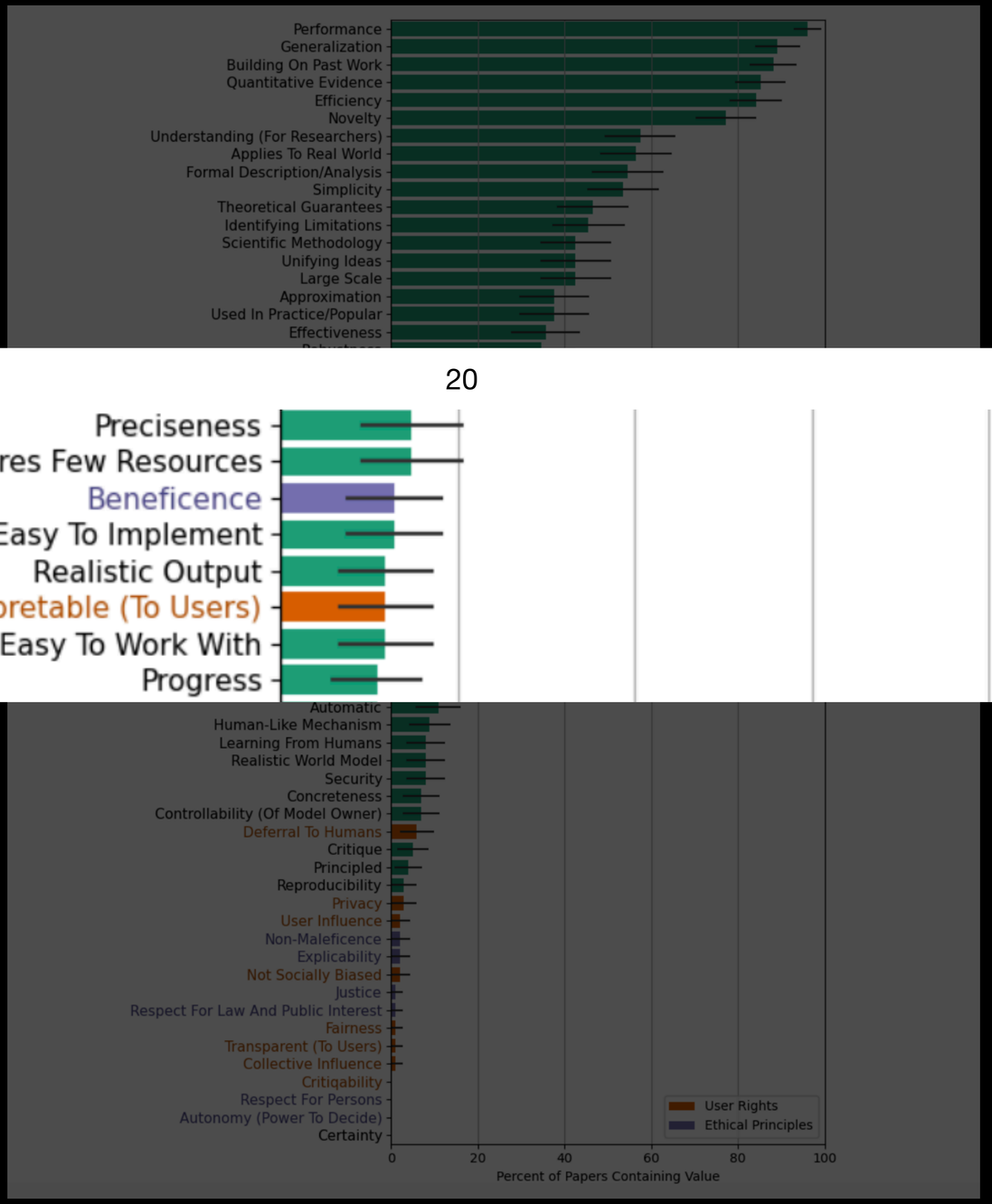
• https://doi.org/10.1145/3531146.3533083

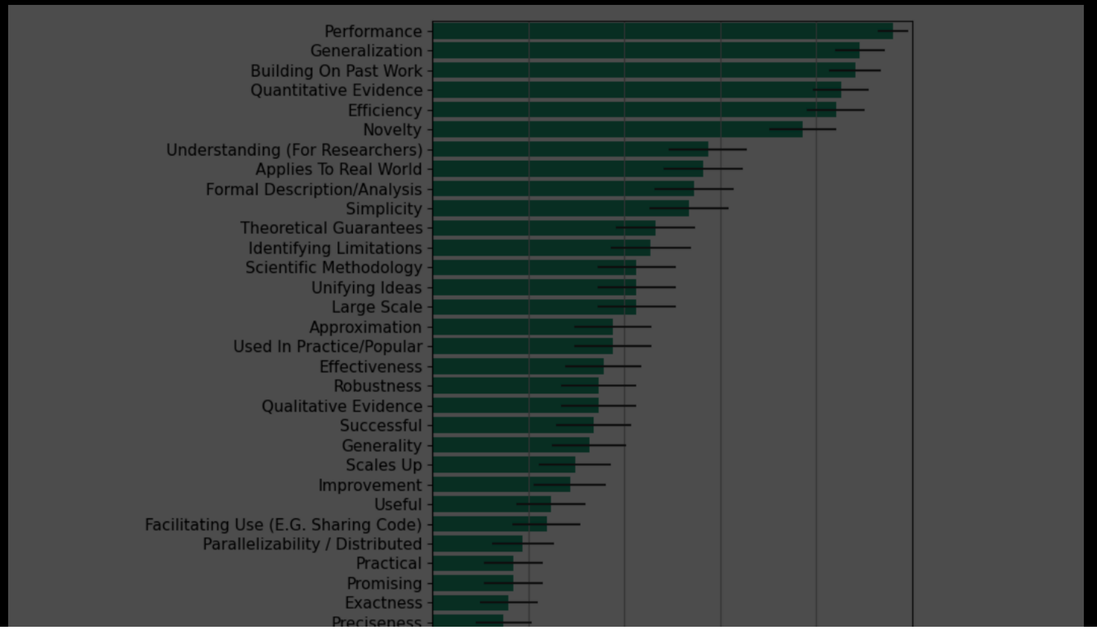
Published: 20 June 2022 Publication History


Performance
Generalization
Building On Past Work
Quantitative Evidence
Efficiency
Novelty
Understanding (For Researchers)
Applies To Real World
Formal Description/Analysis
Simplicity
Theoretical Guarantees
Identifying Limitations
Scientific Methodology
Unifying Ideas
Large Scale
Approximation
Used In Practice/Popular
Effectiveness
Robustness
Qualitative Evidence
Successful
Generality
Scales Up
Improvement
Useful
Facilitating Use (E.G. Sharing Code)
Parallelizability / Distributed
Practical
Promising
Exactness
Preciseness
Requires Few Resources
Beneficence
Easy To Implement
Realistic Output
Interpretable (To Users)
Easy To Work With
Progress
Automatic
Human-Like Mechanism
Learning From Humans
Realistic World Model
Security
Concreteness
Controllability (Of Model Owner)
Deferral To Humans
Critique
Principled
Reproducibility
Privacy
User Influence
Non-Maleficence
Explicability
Not Socially Biased
Justice
Respect For Law And Public Interest
Fairness
Transparent (To Users)
Collective Influence
Critiqability
Respect For Persons
Autonomy (Power To Decide)
Certainty
User Rights
Ethical Principles
0
20
40
60
80
100
Percent of Papers Containing Value

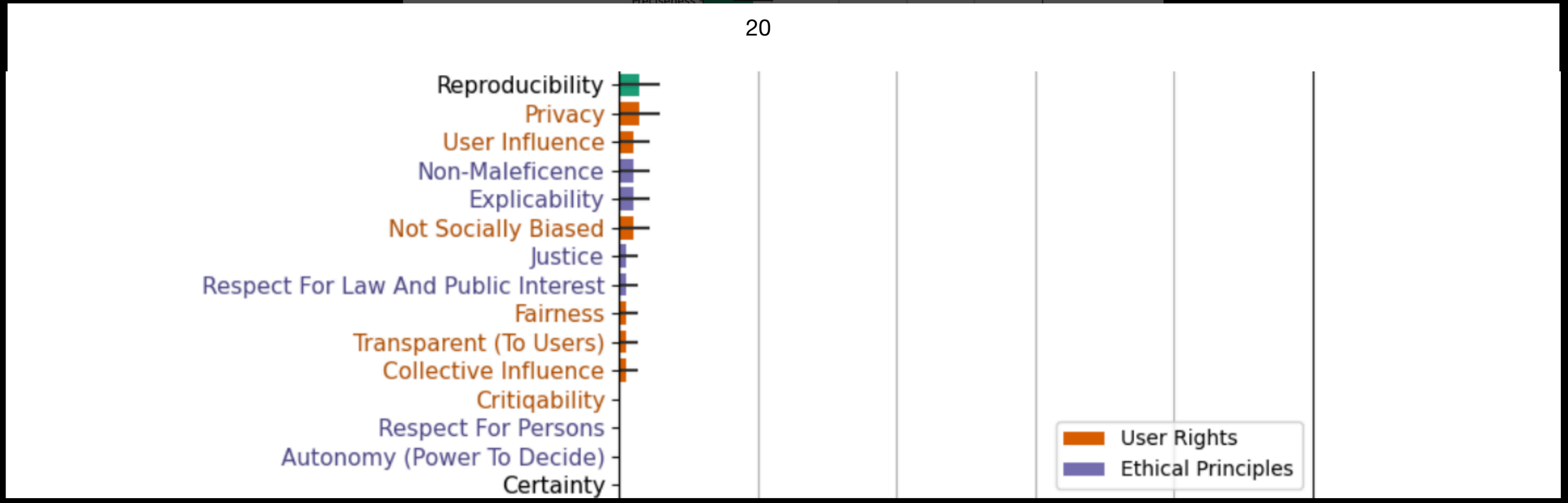

Performance
Generalization
Building On Past Work
Quantitative Evidence
Efficiency
Novelty
Understanding (For Researchers)
Applies To Real World
Formal Description/Analysis
Simplicity
Theoretical Guarantees
Identifying Limitations


Parallelizability / Distributed
Practical
Promising
Exactness
Preciseness
Requires Few Resources
Beneficence
Easy To Implement
Realistic Output
Interpretable (To Users)
Easy To Work With
Progress
Automatic
Human-Like Mechanism
Learning From Humans
Realistic World Model
Security
Concreteness
Controllability (Of Model Owner)
Deferral To Humans
Critique
Principled
Reproducibility
Privacy
User Influence
Non-Maleficence
Explicability
Not Socially Biased
Justice
Respect For Law And Public Interest
Fairness
Transparent (To Users)
Collective Influence
Critiqability
Respect For Persons
Autonomy (Power To Decide)
Certainty
User Rights
Ethical Principles
0
20
40
60
80
100
Percent of Papers Containing Value

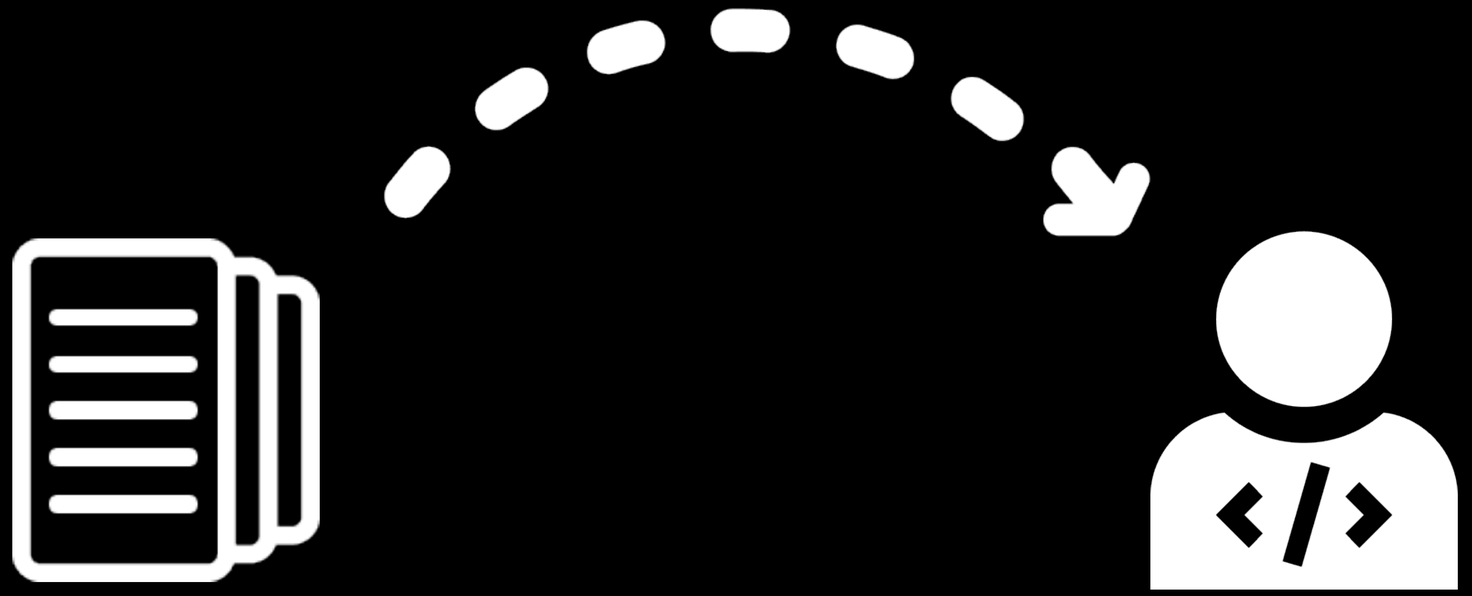

Performance
Generalization
Building On Past Work
Quantitative Evidence
Efficiency
Novelty
Understanding (For Researchers)
Applies To Real World
Formal Description/Analysis
Simplicity
Theoretical Guarantees
Identifying Limitations
Scientific Methodology
Unifying Ideas
Large Scale
Approximation
Used In Practice/Popular
Effectiveness
20
Preciseness
Requires Few Resources
Beneficence
Easy To Implement
Realistic Output
Interpretable (To Users)
Easy To Work With
Progress
Automatic
Human-Like Mechanism
Learning From Humans
Realistic World Model
Security
Concreteness
Controllability (Of Model Owner)
Deferral To Humans
Critique
Principled
Reproducibility
Privacy
User Influence
Non-Maleficence
Explicability
Not Socially Biased
Justice
Respect For Law And Public Interest
Fairness
Transparent (To Users)
Collective Influence
Critiqability
Respect For Persons
Autonomy (Power To Decide)
Certainty
User Rights
Ethical Principles
0
20
40
60
80
100
Percent of Papers Containing Value


Performance
Generalization
Building On Past Work
Quantitative Evidence
Efficiency
Novelty
Understanding (For Researchers)
Applies To Real World
Formal Description/Analysis
Simplicity
Theoretical Guarantees
Identifying Limitations
Scientific Methodology
Unifying Ideas
Large Scale
Approximation
Used In Practice/Popular
Effectiveness
Robustness
Qualitative Evidence
Successful
Generality
Scales Up
Improvement
Useful
Facilitating Use (E.G. Sharing Code)
Parallelizability / Distributed
Practical
Promising
Exactness
Preciseness

20


Reproducibility
Privacy
User Influence
Non-Maleficence
Explicability
Not Socially Biased
Justice
Respect For Law And Public Interest
Fairness
Transparent (To Users)
Collective Influence
Critiqability
Respect For Persons
Autonomy (Power To Decide)
Certainty
User Rights
Ethical Principles

Outside of our data, what are we actually working on?



The study of bias has expanded from just the dataset to also include the developer

The reality is now more than ever, the data of LLMs is out of our hands

- Foundational models
- Closed-source

The reality is now more than ever, the data of LLMs is out of our hands

- Foundational models
- Closed-source

While data issues persist (copyrighted material, artist style)

The reality is now more than ever, the data of LLMs is out of our hands

- Foundational models
- Closed-source

While data issues persist (copyrighted material, artist style), it matters what we choose to now build

[Submitted on 28 May 2020 (v1), last revised 29 May 2020 (this version, v2)]

Language (Technology) is Power: A Critical Survey of "Bias" in NLP

Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach

We survey 146 papers analyzing "bias" in NLP systems, finding that their motivations are often vague, inconsistent, and lacking in normative reasoning, despite the fact that analyzing "bias" is an inherently normative process. We further find that these papers' proposed quantitative techniques for measuring or mitigating "bias" are poorly matched to their motivations and do not engage with the relevant literature outside of NLP. Based on these findings, we describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems. These recommendations rest on a greater recognition of the relationships between language and social hierarchies, encouraging researchers and practitioners to articulate their conceptualizations of "bias"---i.e., what kinds of system behaviors are harmful, in what ways, to whom, and why, as well as the normative reasoning underlying these statements---and to center work around the lived experiences of members of communities affected by NLP systems, while interrogating and reimagining the power relations between technologists and such communities.

We are under-defining our motivations for the values we are choosing to optimize for — even when the value seems objectively good

[Submitted on 28 May 2020 (v1), last revised 29 May 2020 (this version, v2)]

Language (Technology) is Power: A Critical Survey of "Bias" in NLP

Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach

We survey 146 papers analyzing "bias" in NLP systems, finding that their motivations are often vague, inconsistent, and lacking in normative reasoning, despite the fact that analyzing "bias" is an inherently normative process. We further find that these papers' proposed quantitative techniques for measuring or mitigating "bias" are poorly matched to their motivations and do not engage with the relevant literature outside of NLP. Based on these findings, we describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems. These recommendations rest on a greater recognition of the relationships between language and social hierarchies, encouraging researchers and practitioners to articulate their conceptualizations of "bias"---i.e., what kinds of system behaviors are harmful, in what ways, to whom, and why, as well as the normative reasoning underlying these statements---and to center work around the lived experiences of members of communities affected by NLP systems, while interrogating and reimagining the power relations between technologists and such communities.

We are under-defining our motivations for the values we are choosing to optimize for — even when the value seems objectively good

We are short-cutting justification

Let’s imagine we are writing an LLM to assist with grant writing for non-profits.

Let’s imagine we are writing an LLM to assist with grant writing for non-profits.

This tool will check for grammar.

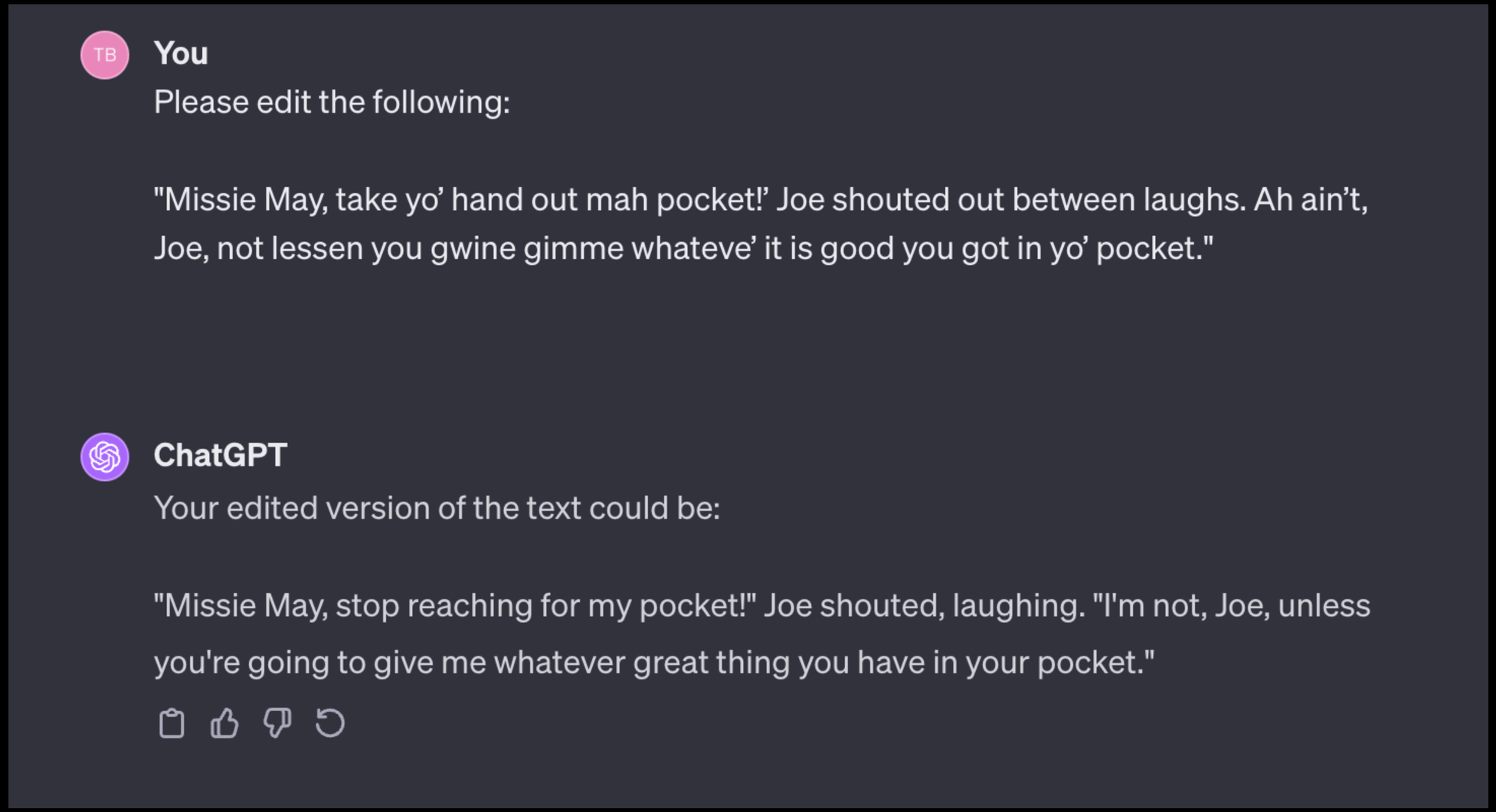
Let’s imagine we are writing an LLM to assist with grant writing for non-profits.

This tool will check for grammar.

But what is “grammar”?

"Even after decades of sociolinguistic efforts to legitimize African-American Vernacular English (AAVE), it continues to be viewed **as "bad" English** and its speakers continue to be viewed as linguistically inadequate—a view called the deficit perspective (Alim et al., 2016; Rosa and Flores, 2017). This perspective persists despite demonstrations that **AAVE is rule-bound and grammatical** (Mufwene et al., 1998; Green, 2002)"

[Blodgett 2020]



[Author, 2023]

[Zora Neale Hurston, "The Gilded Six-Bits" 1933]

TB **You**

Please edit the following:

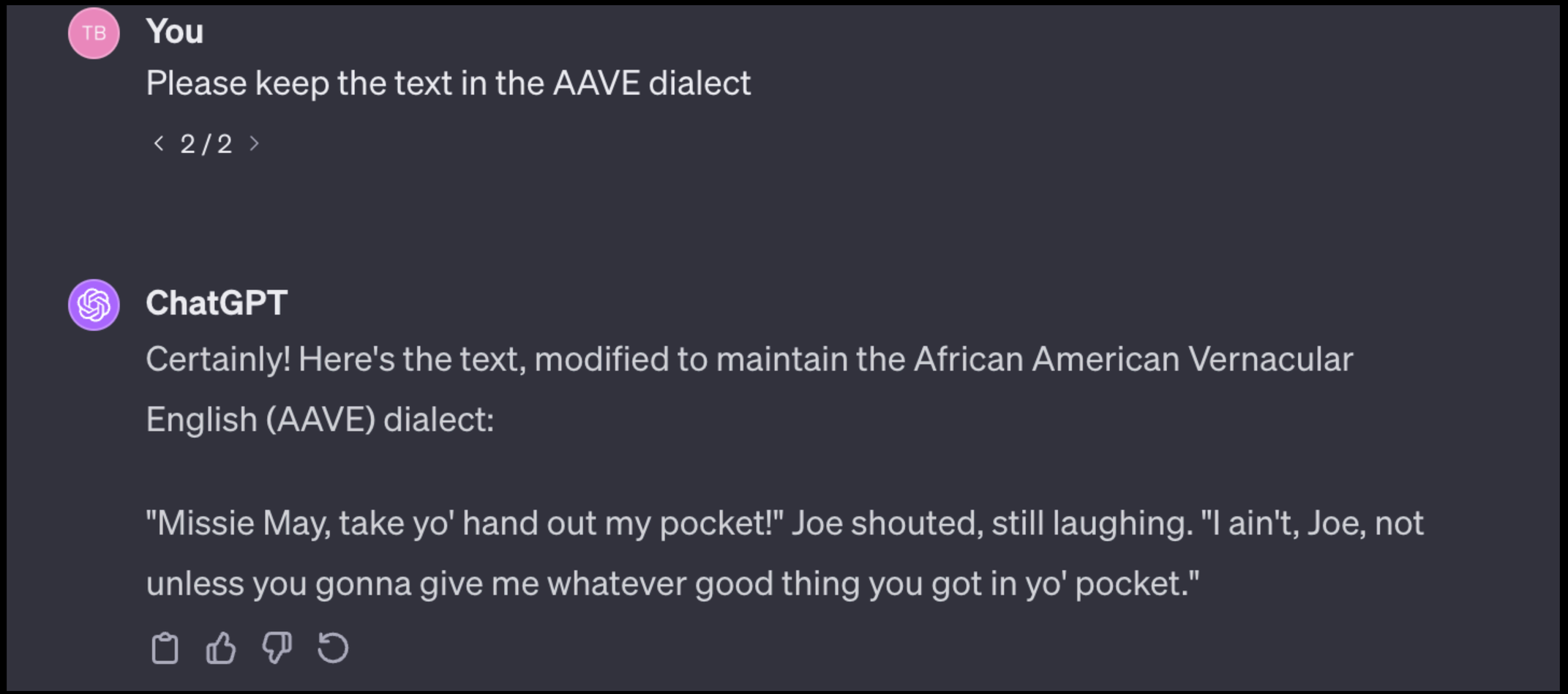
"Missie May, take yo' hand out mah pocket!' Joe shouted out between laughs. Ah ain't, Joe, not lessen you gwine gimme whateve' it is good you got in yo' pocket."

TB **You**

Please edit the following:

"Missie May, stop reaching for my pocket!" Joe shouted, laughing. "I'm not, Joe, unless you're going to give me whatever great thing you have in your pocket."

[Author, 2023]


TB
You
Please keep the text in the AAVE dialect
2 / 2
ChatGPT
Certainly! Here's the text, modified to maintain the African American Vernacular English (AAVE) dialect:
"Missie May, take yo' hand out my pocket!" Joe shouted, still laughing. "I ain't, Joe, not unless you gonna give me whatever good thing you got in yo' pocket."

[Author, 2023]

By enforcing American Standard English (ASE) grammar as the default, who are we building for? Who are we excluding?

“In **re-stigmatizing** AAVE, [language tech] reproduces language ideologies in which AAVE is viewed as ungrammatical, uneducated, and offensive.”

[Blodgett 2020]

The reality is though that ASE grammar is often seen as the only “acceptable” English.

It is likely that a grant app written outside of ASE might not be accepted. Allocational harms do exist for AAVE speakers.

The reality is though that ASE grammar is often seen as the only “acceptable” English.

It is likely that a grant app written outside of ASE might not be accepted. Allocational harms do exist for AAVE speakers.

But our tool should make explicit that it will edit into ASE. Or instead, it should ask the user what editing help they would prefer.

Our AI grant writing tool can enforce ASE grammar, but it should not do so without reflection on this decision and giving agency to the user

Let’s imagine we are working on a machine translation LLM.

Let’s imagine we are working on a machine translation LLM.

Let’s include as many languages as possible.

“So for example, making languages available for translation might also make them and their speakers available for ***increased surveillance***, or might mean making a community's cultural and linguistic resources ***available to the wider world in ways that they might not want***. Or, depending on who develops the data might means that somebody who's not the community ***profits*** from these cultural resources.”

[Blodgett, “The Good Robot” podcast, 2022]

“So even things that we kind of default assume are good, like inclusion are surprisingly fraught.”

[Blodgett, “The Good Robot” podcast, 2022]

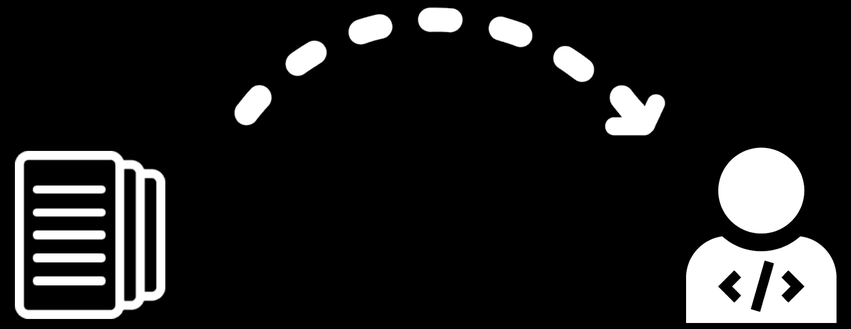
WORLD

Indigenous groups fear culture distortion as AI learns their languages

"Data is like our land and natural resources," said Karaitiana Taiuru, a Maori ethicist and an honorary academic at the University of Auckland.

"If Indigenous peoples ***don't have sovereignty of their own data***, they will simply be re-colonized in this information society."

Our translation tool should include multiple world languages, esp. those at risk of dying off. However, we cannot just add in all languages without reflection.



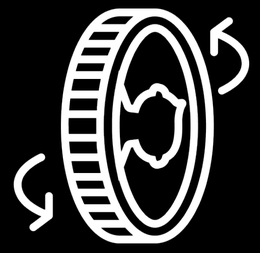
As datasets have been standardized, we now have *more* opportunity and *more* agency to consider the values we build for

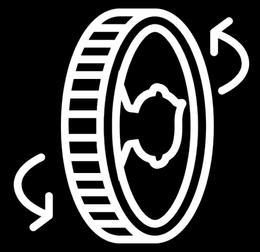
“Good” isn’t good enough

Ben Green
Harvard University
bgreen@g.harvard.edu

Abstract

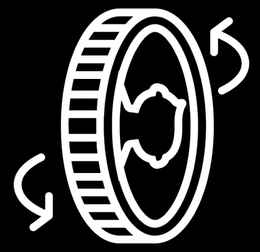
Despite widespread enthusiasm among computer scientists to contribute to “social good,” the field’s efforts to promote good lack a rigorous foundation in politics or social change. There is limited discourse regarding what “good” actually entails, and instead a reliance on vague notions of what aspects of society are good or bad. Moreover, the field rarely considers the types of social change that result from algorithmic interventions, instead following a “greedy algorithm” approach of pursuing technology-centric incremental reform at all points. In order to reason well about doing good, computer scientists must reflexively evaluate their normative commitments, consider the long-term impacts of technological interventions, evaluate algorithmic interventions against alternative reforms, and no longer prioritize technical considerations as superior to other forms of knowledge.





“Enhance police accountability and promote non-punitive alternatives to incarceration”

[Green 2019]



“Enhance police accountability and promote non-punitive alternatives to incarceration”

“while other work uses data to predict and classify crimes to aid police investigations”

[Green 2019]

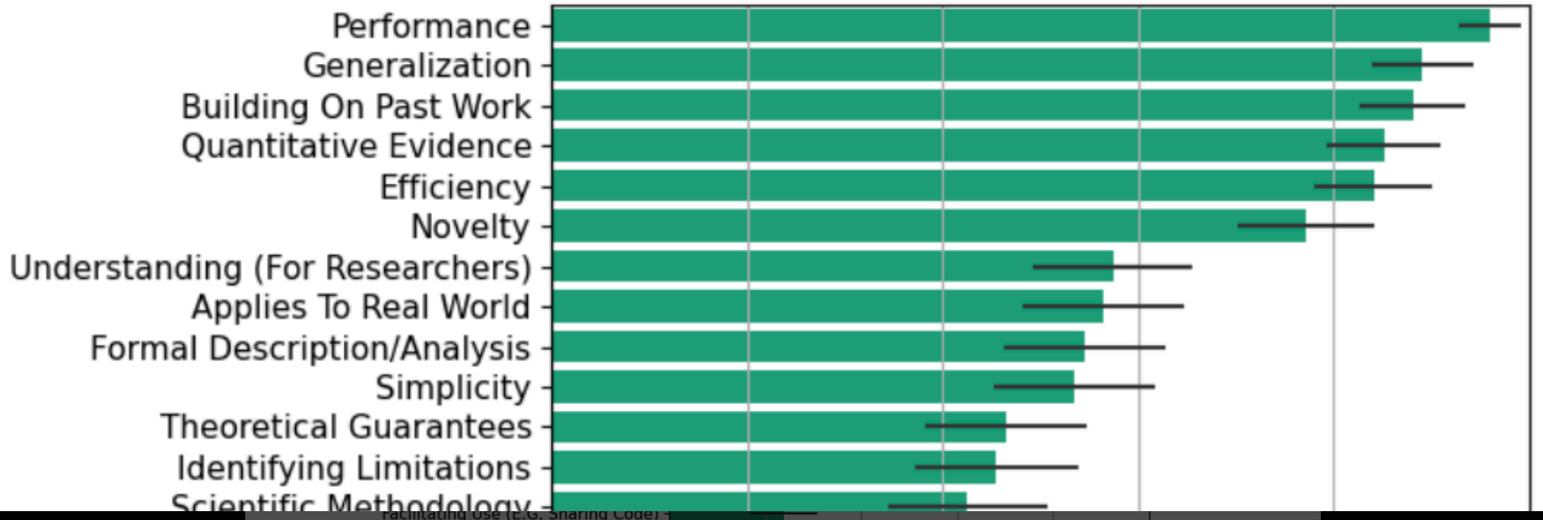
How can two opposed projects both be justifying themselves through “social good”?

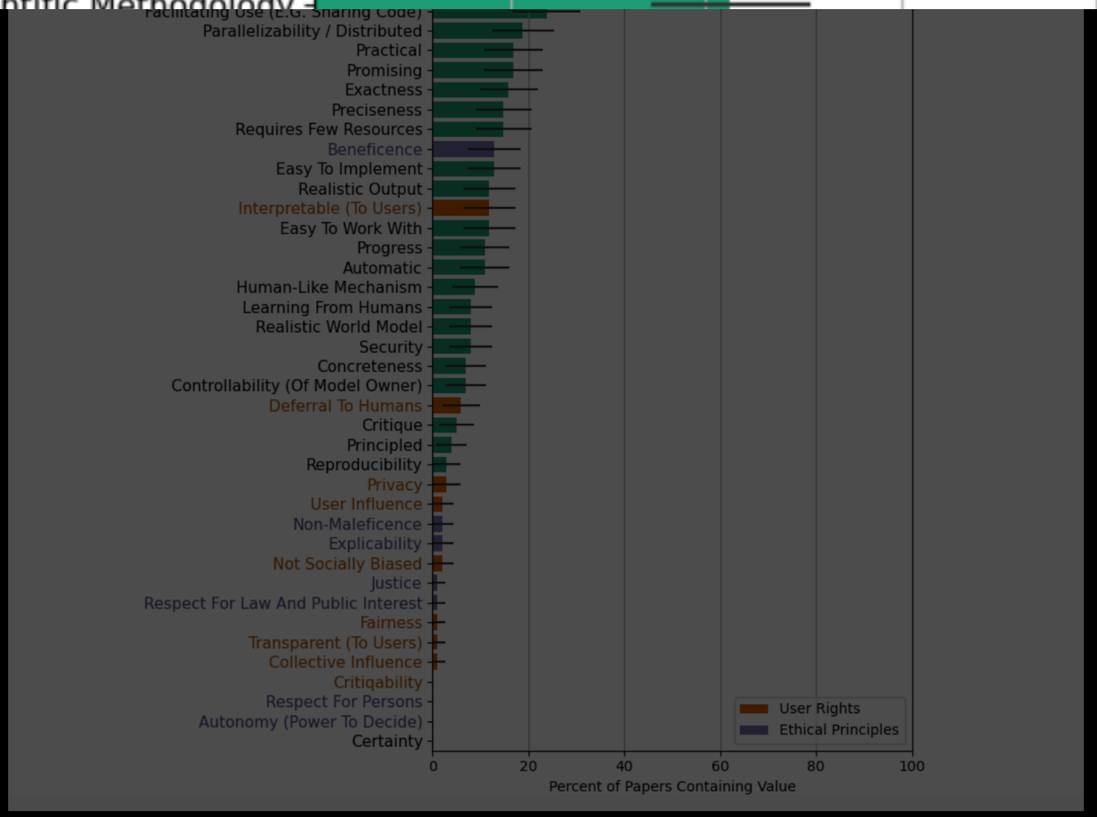
How can two opposed projects both be justifying themselves through “social good”?

This is the slipperiness of the terms.

"Rather than referring to ‘social good,’ computer scientists should ***more explicitly consider and articulate the normative commitments*** behind their work (whatever they may be)”

[Green 2019]


Performance
Generalization
Building On Past Work
Quantitative Evidence
Efficiency
Novelty
Understanding (For Researchers)
Applies To Real World
Formal Description/Analysis
Simplicity
Theoretical Guarantees
Identifying Limitations


Parallelizability / Distributed
Practical
Promising
Exactness
Preciseness
Requires Few Resources
Beneficence
Easy To Implement
Realistic Output
Interpretable (To Users)
Easy To Work With
Progress
Automatic
Human-Like Mechanism
Learning From Humans
Realistic World Model
Security
Concreteness
Controllability (Of Model Owner)
Deferral To Humans
Critique
Principled
Reproducibility
Privacy
User Influence
Non-Maleficence
Explicability
Not Socially Biased
Justice
Respect For Law And Public Interest
Fairness
Transparent (To Users)
Collective Influence
Critiqability
Respect For Persons
Autonomy (Power To Decide)
Certainty
User Rights
Ethical Principles
0
20
40
60
80
100
Percent of Papers Containing Value

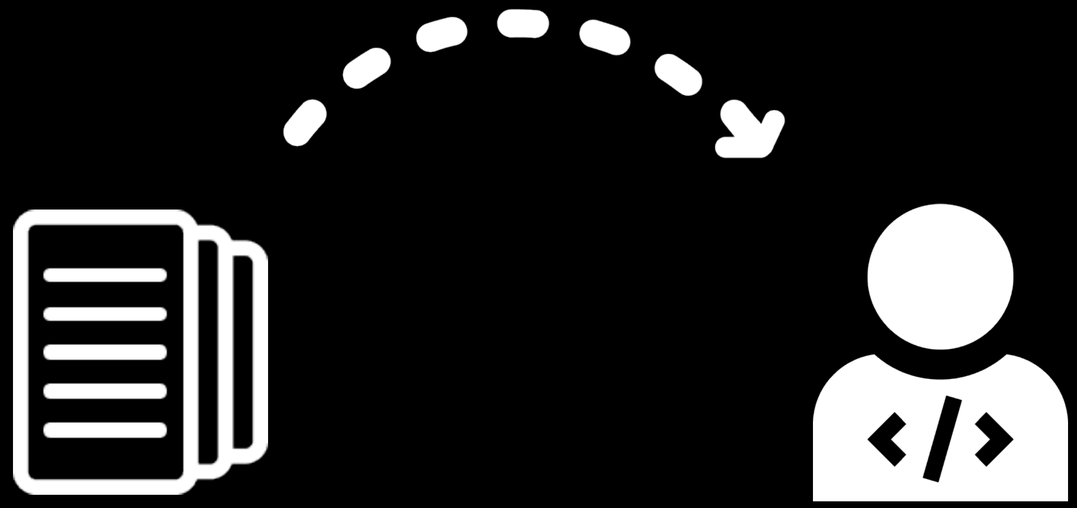
THE PROBLEM WITH BIAS

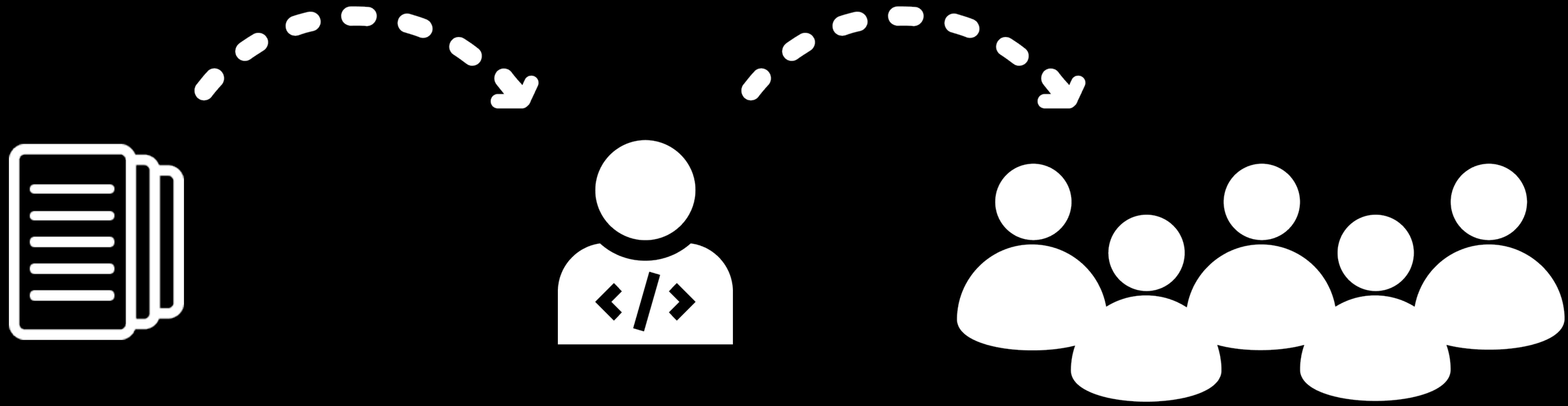
Allocation	Representation
Immediate	Long term
Easily quantifiable	Difficult to formalize
Discrete	Diffuse
Transactional	Cultural

[Crawford, “The Trouble with Bias,” 2017]

What seem to be objectively good values inclusion, grammaticality, social good

What seem to be objectively good values



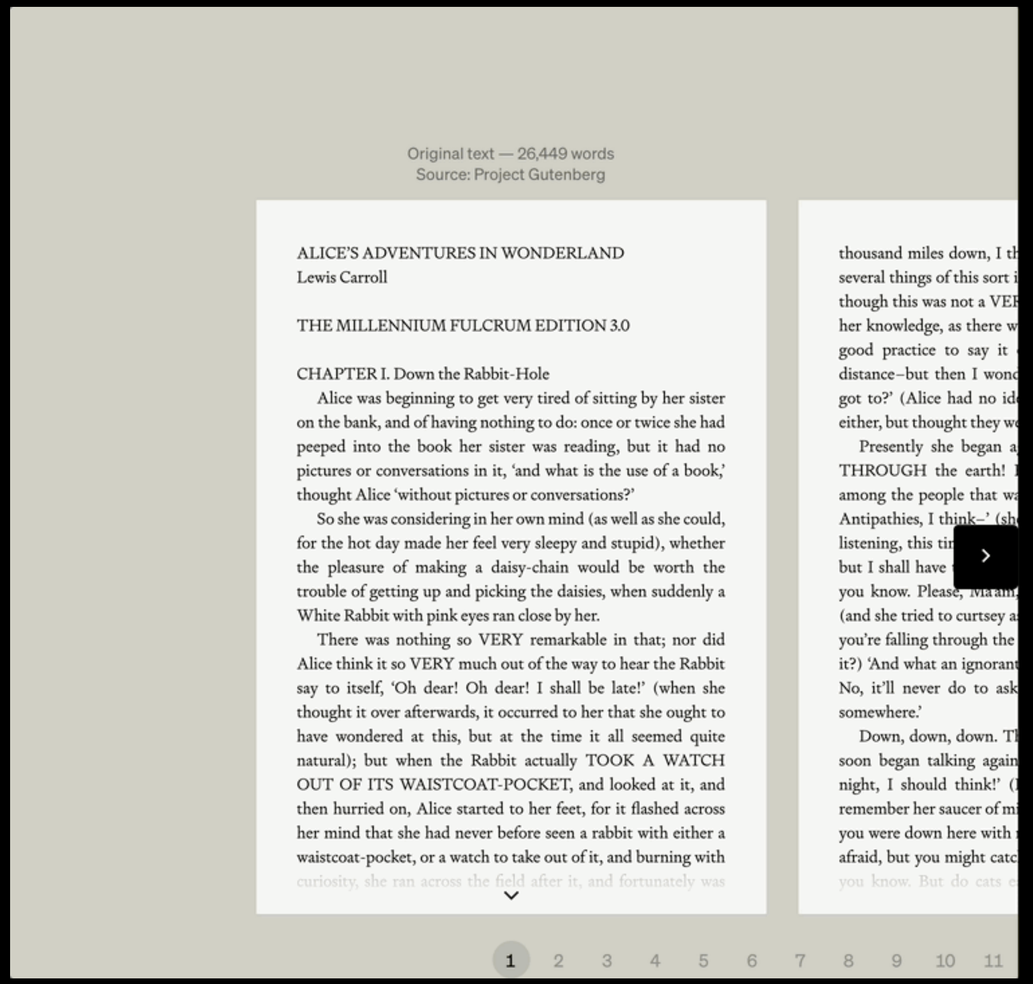


RLHF

Reinforcement Learning from Human Feedback

Research

Summarizing books with human feedback



Recursive task decomposition — they summarized books to chapters to sections to sentences

At every stage, they asked for human feedback on the quality of the summary

But efficient summarization seems fraught when the goal of literature is arguably ***friction, escape, slowness***

Asking which summary users prefer is not the same thing as justifying the need for literature summary at all

In their instructions for human labelers, OpenAI informs them that “gratuitously including small details is generally penalized.”

But what is literature if not a set of small details that come together?

[“Recursively summarizing books with human feedback.” 2021]

Depth 2

Daisy tells the narrator that she's cynical about everything and that she said something when her daughter was born.

At home, the narrator sees Mr. Gatsby standing outside, looking at the stars.

The narrator recalls the first time he met Tom's mistress. He traveled to New York with Tom on a train and Tom insisted he meet her. Mrs. Wilson, Tom's mistress, buys a dog and invites everyone to her apartment. The party gets lively as more guests arrive.

Catherine, Mrs. Wilson's sister, sits next to the narrator on the couch and asks if he knows Gatsby, who lives next door to him.

Copied directly

Depth 3

The telephone rings inside, and the subject of the horses is not mentioned again. Tom and Miss Baker return to the library, while Daisy and the narrator go to the porch.

Daisy tells the narrator that she has had a very bad time, and that she is cynical about everything. She then tells him that she said something when her daughter was born. She says that she woke up from the ether with an abandoned feeling, and asked the nurse if the baby was a boy or a girl. She was told it was a girl.

Depth 4 (source text) LICENSE

hardy scepticism, was able utterly to put this fifth guestâ□□s shrill metallic urgency out of mind. To a certain temperament the situation might have seemed intriguingâ□□my own instinct was to telephone immediately for the police.

Mood
Setting

The horses, needless to say, were not mentioned again. Tom and Miss Baker, with several feet of twilight between them, strolled back into the library, as if to a vigil beside a perfectly tangible body, while,

Copied directly
Imagery
Simile

trying to look pleasantly interested and a little deaf, I followed Daisy around a chain of connecting verandas to the porch in front. In its deep gloom we sat down side by side on a wicker settee.

Imagery
Setting
Personification

Daisy took her face in her hands as if feeling its lovely shape, and her eyes moved gradually out into the velvet dusk. I saw that turbulent emotions possessed her, so I asked what I thought would be some sedative questions about her little girl.

Imagery
Turn of phrase
Mood

â□□We donâ□□t know each other very well, Nick,â□□ she said suddenly. â□□Even if we are cousins. You didnâ□□t come to my wedding.â□□

â□□I wasnâ□□t back from the war.â□□

â□□Thatâ□□s true.â□□ She hesitated.

Dialogue
Relationship
Setting

â□□Well, Iâ□□ve had a very bad time, Nick, and Iâ□□m pretty cynical about everything.â□□

Copied directly
Copied directly

Evidently she had reason to be. I waited but she didnâ□□t say any more, and after a moment I returned rather feebly to the subject of her daughter.

â□□I suppose she talks, andâ□□eats, and everything.â□□

â□□Oh, yes.â□□ She looked at me absently. â□□Listen, Nick; let me tell you what I said when she was born. Would you like to hear?â□□

â□□Very much.â□□

â□□Itâ□□ll show you how Iâ□□ve gotten to feel aboutâ□□things. Well, she was less than an hour old and Tom was God knows where. I woke up out of the ether with an utterly abandoned feeling, and asked the nurse right away if it was a boy or a girl. She told me it was a girl, and so I

Copied directly

It is bizarre that human feedback was being used to remove the “literature” from literature… and that this was considered aligned because of user preferences of which summary is better

The method is aligned, but the goal is not.

This focus on scaling feedback is part of the *un*alignment: ***“Literary achievement is indifferent to the mindset of efficiency,”*** and yet here literature is being used to advance summarization efficiency

[Michele Elam, “Poetry Will Not Optimize; or, What Is Literature to AI?.” 2023]

Which response do you prefer?

Your choice will help make ChatGPT better.

Response 1

Certainly! Here's a more polished message:

"Congratulations @Fanilo on surpassing the 5k subscriber milestone on YouTube yesterday! 🎉

https://twitter.com/andfanilo/status/1705492081202880993"

Response 2

Certainly! Here's an enhanced version:

"🎉 Kudos to @Fanilo for surpassing the 5k subscriber milestone on YouTube! Such an impressive achievement. Keep soaring! 🚀

https://twitter.com/andfanilo/status/1705492081202880993"

Which response do you prefer?

Your choice will help make ChatGPT better.

Response 1

Certainly! Here's a more polished message:

"Congratulations @Fanilo on surpassing the 5k subscriber milestone on YouTube yesterday! 🎉

https://twitter.com/andfanilo/status/1705492081202880993"

Response 2

Certainly! Here's an enhanced version:

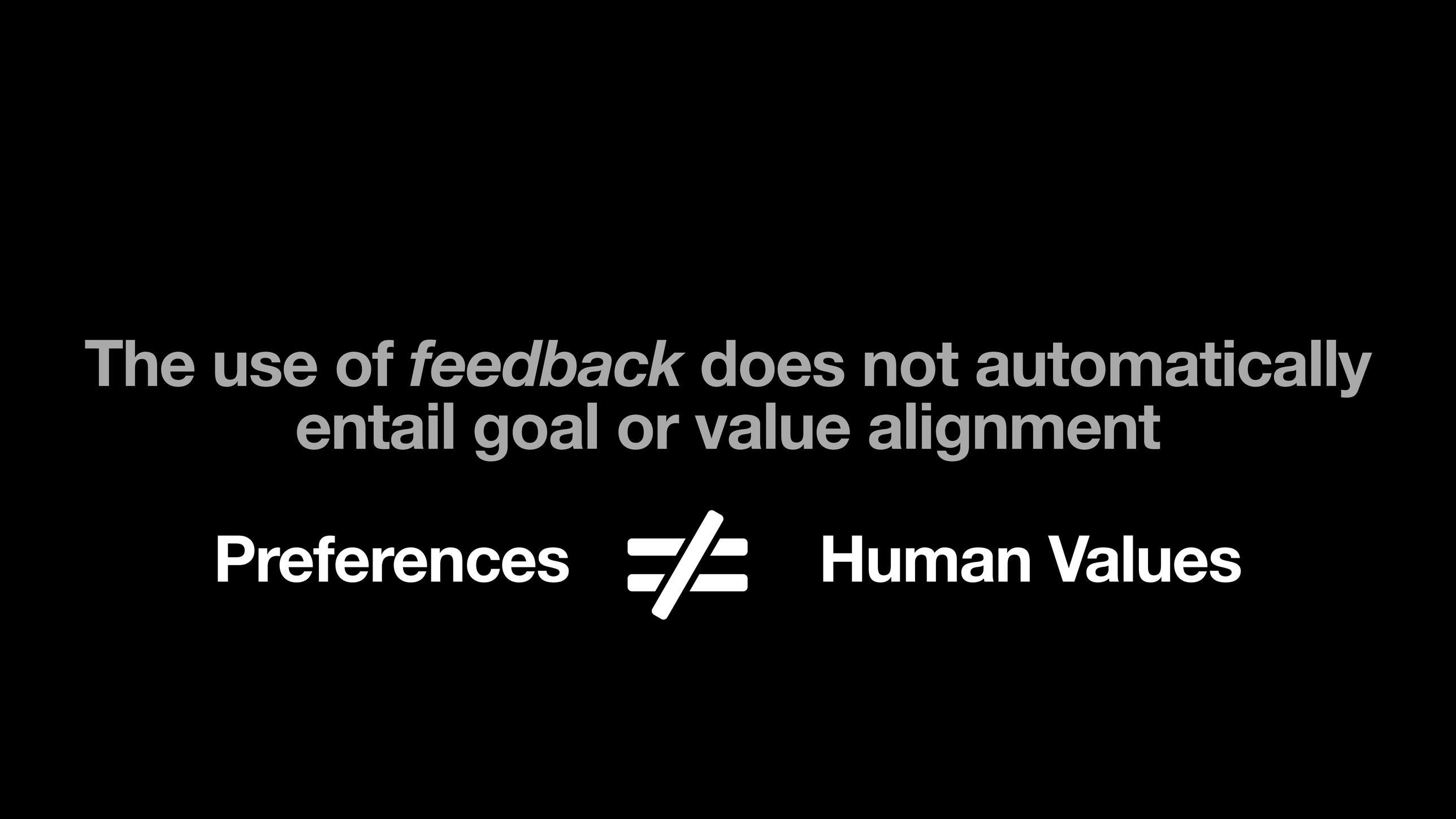
"🎉 Kudos to @Fanilo for surpassing the 5k subscriber milestone on YouTube! Such an impressive achievement. Keep soaring! 🚀

https://twitter.com/andfanilo/status/1705492081202880993"

Alignment can be powerful and beautiful…
but it’s not perfect

The use of *feedback* does not automatically entail goal or value alignment

The use of *feedback* does not automatically entail goal or value alignment



Power to the People? Opportunities and Challenges for Participatory AI

Abeba Birhane
Mozilla Foundation & University College Dublin
Ireland
abeba.birhane@ucdconnect.ie

William Isaac
DeepMind
UK
williamis@deepmind.com

Vinodkumar Prabhakaran
Google
USA
vinodkpg@google.com

Mark Díaz
Google
USA
markdiaz@google.com

Madeleine Clare Elish
Google
USA
mcelish@google.com

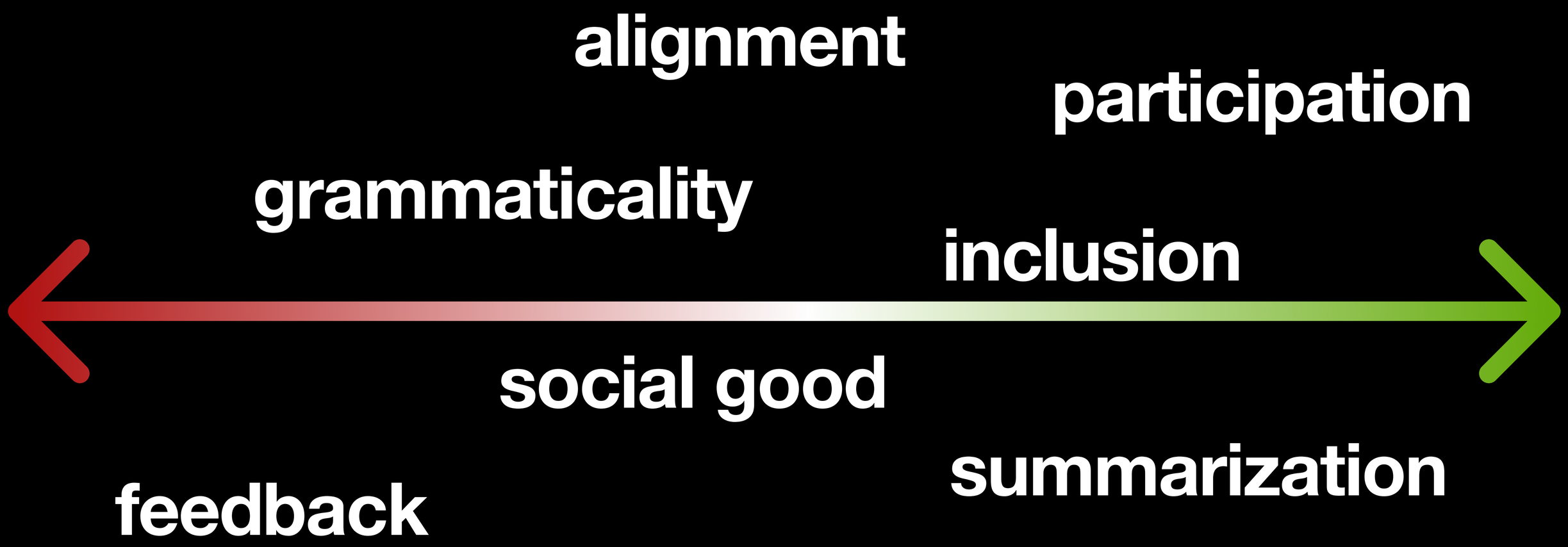
Iason Gabriel
DeepMind
UK
iason@deepmind.com

Shakir Mohamed
DeepMind
UK
shakir@deepmind.com

"Yet, caution has been raised about the possibility of ***'participation-washing'***, where efforts are mischaracterized under the banner of participation, are weakly-executed, or co-opt the voice of participants to ***achieve predetermined aims***"

[Birhane 2022]


grammaticality
inclusion
social good


alignment
participation
grammaticality
inclusion
social good
summarization
feedback

“So even things that we kind of default assume are good, like inclusion are surprisingly fraught.”

[Blodgett, “The Good Robot” podcast, 2022]

So....

My goal is not to scare, nor to stop LLM work,
nor to shut-down human alignment research

I want these things to work

From ProPublica to Buolamwini and Gebru to Crawford to Blodgett to Green…

AI researchers want to see the benefits of such tech actualized.

From ProPublica to Buolamwini and Gebru to Crawford to Blodgett to Green…

AI researchers want to see the benefits of such tech actualized.

But it cannot be done if we are on the side-lines about the values and assumptions we are building into our own tools

As your wrap-up your final projects…

I want you to engage with users, to try and be aligned.
I want you to have productive, performing tools.

As your wrap-up your final projects…

I want you to engage with users, to try and be aligned.
I want you to have productive, performing tools.

But I also want you to be aware that all decisions we make as developers have pros and cons.

What values are your optimizing for? Why do you think those are justified? What do those values exclude?

I encourage you to be explicit about the embedded assumptions, norms, and values

for yourselves, for your users, for 3rd party auditors

~I left a lot out~

Interpretability, ghost work, evaluation, data privacy

Tommy’s Course Recommendations on Value-Based AI and LLMs

Technical Ethics

CS281: Ethics of AI (S)

AI Analysis of Humanities

ENGLISH 184F: Literary Text Mining 2: Studies in Cultural Analytics (W)

Understanding Ethics and Politics of Language/Grammar

AFRICAAM 389C: Black Digital Cultures from BlackPlanet to AI (W)
ENGLISH 362C: Language Politics and the Literary Imagination in Africa (S)

Unpacking Latent Cultural Values in Science

ANTHRO 120H: Intro to Medical Humanities (W)
AFRICAAM 151: Ethical STEM: Race, Justice, and Embodied Practice (W)
JAPAN 151B: The Nature of Knowledge: Science and Literature in East Asia (W)

CS224V: Ethics and Policy Lecture

Value-Based LLMs

How Bias, Alignment, and your values come together

Tommy Bruzzese

Wed Nov 29th, 2023