



Ethical AI & product development @ Wikimedia

Jonathan T. Morgan

Workshop on AI at Wikimedia • Berkman Klein Center • 25 June 2019

meta.wikimedia.org/wiki/Research:Ethical_AI

Good morning. My name is Jonathan Morgan. I'm a design researcher on the central Research team at Wikimedia. Today I'm going to present to you some ongoing work focused on improving the way Wikimedia develops software tools, products, and resources that utilize machine learning.

Goal: Improve *AI product* development

Help the Wikimedia Foundation...

- **Leverage AI capabilities** to achieve our strategic goals
- **Align our decisions with our values:** e.g. what we build, how we build it, how we define success (and failure)
- **Avoid unintended consequences** that undermine our goals, values, the quality of Wikipedia, the efforts of its contributors, or the trust of its readers
- **Make informed decisions** including about difficult tradeoffs
- **Continue to learn and evolve**

meta.wikimedia.org/wiki/Research:Ethical_AI

The overall goal of this project is to understand how Wikimedia can use ML or "AI" technologies to further strategic goals like making high quality information available in hundreds of languages, while avoiding unintended consequences that could undermine those goals. My intent is to help our organization leverage the benefits of machine learning technologies, while assuring that the product decisions we make are aligned with our values, that we make well-informed choices and tradeoffs, and that we nurture a capacity to continuously to learn, improve, and evolve our processes to keep pace with the technological landscape, new research, emerging industry best practices, and regulations.

https://meta.wikimedia.org/wiki/Research:Ethical_AI

Guiding question

What would a *Minimum Viable Process* for ethical AI at Wikimedia look like?

Given our...

- free-culture mission and open-source ethos
- organizational processes and practices
- relatively small size
- shared governance of Wikipedia



The overall guiding question for the project is "what would a minimum viable process" for ethical AI product development look like in an organization like Wikimedia, given the features of our organization and our movement that set us apart from other technology companies. Features such as our free culture ethos and mission, our current processes and practices for developing software, the relatively small size of our product org, and our shared governance over Wikipedia with the volunteer communities that create and curate the content.

Wikimedia's AI Products

- **ML-driven applications** software and hardware with user interfaces
- **Machine learning models** & their supporting code and documentation
- **ML platforms** public APIs; hosting & integration infrastructure for models
- **Public datasets** used for training ML models



These features of our organization and movement directly shape the kinds of products we build and release. For example, in most technology companies the primary kind of "AI product" is some piece of software or hardware that has a user interface, a freestanding product like a Amazon Echo or a product feature like Facebook's newsfeed. Many technology companies will only "release" these kinds of AI products, with the possible addition of ML platforms if they're in a service business. The models themselves and the underlying data are closely guarded intellectual property. But Wikimedia is as an open source software and open data company, so all of the elements in the AI technology stack are public, and are therefore AI products in their own right--from the source code for the machine learning models we develop, the public web APIs and machine learning as a service infrastructure that allow people to develop tools and perform research with those models, and the datasets we use to train them. Even the purpose-built tools and interfaces we use to label the training data.

Because they're all public, each of these types of AI product can be used for a variety of intended and unintended purposes, separately or in combination with other AI products, within or outside of our particular technological ecosystem. These combinations create a truly dizzying number of ethical implications.

Ethical AI principles

1. Fair
2. Transparent
3. Accountable

WIKIMEDIA
FOUNDATION

The core ethical AI principles of fairness, transparency, and accountability give us a general idea of what we need to shoot for, and what's at stake different phases of product development: planning, development, deployment, and maintenance. But applying these principles to a specific product development contexts at a particular company can be a challenge.

Ethical AI principles

1. Fair
2. Transparent
3. Accountable

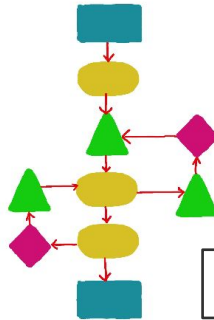


WIKIMEDIA
FOUNDATION

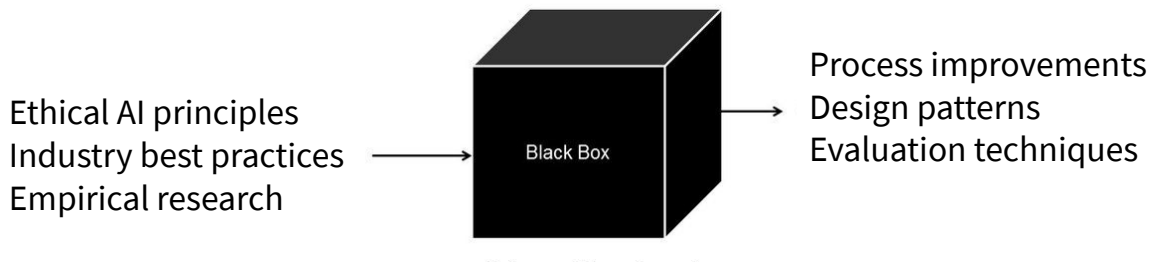
There's a great deal of the guidance on how to "do" ethical AI out there, most of it developed over the past 3-5 years. In my experience, it tends to be either too broad or too specific to be actionable in an industry context: the main types I've seen are high-level policy and position statements that articulate what these principles mean and why they are important, retrospective case studies of AI products gone wrong, formal mathematical models of fairness, etc, and empirical evaluations of specific models or datasets.

Ethical AI principles

1. Fair
2. Transparent
3. Accountable



What's missing so far is concrete guidance on how to adapt and integrate these principles, cases, models, and findings to an organization's existing product development workflows, synthesize them into design patterns to inform the development of AI product interfaces and documentation, or inform decision-making around prioritization and trade-offs.



So one of the things I've been grappling with over the past year or so is how to distill the wealth of guidance on ethical AI drawn from whatever literature I can find into a suite of process improvements, design patterns, and evaluation methods that work for Wikimedia.

Process proposals: Identify general categories of ethical AI design & development guidance via a literature review

Risk scenarios: Characterize specific risks and trade-offs grounded in Wikimedia's unique AI product context

Ethical & Human-centered AI

Executive summary

AI technologies have the potential to benefit the Wikimedia Movement, but they come with risks. The Wikimedia Foundation has begun to build AI products around these technologies. The emerging domain of *ethical AI* proposes new approaches for addressing the discrimination, disruption, and damage that AI can cause. The established discipline of *human centered design* provides guidance on how to maintain a focus on human needs and well-being throughout product development.

This white paper is intended to help Wikimedia ensure ethical and human-centered outcomes in AI product development. It gives our current and anticipated goals, needs, capacities, and weaknesses. The paper makes two contributions: (1) it introduces a set of risk scenarios intended to define the problem space and promote reflective decision-making; and (2) it generates a set of *process proposals* for improving AI product development. Taken together, these scenarios and proposals can help Wikimedia address anticipated challenges and identify emerging opportunities to leverage AI technologies to further our mission.

This document is oriented towards Wikimedia's 2027 Strategic Direction. It complements the strategic priorities described in the white papers *Knowledge gaps, Knowledge strategy, and Foundation*¹ by Wikimedia Research and *Augmentation*² by Wikimedia Audiences. Background material and additional resources are available on meta.wikimedia.org.

Cite this document as: Jonathan T. Morgan, 2019, *Ethical & Human-Centered AI* - Wikimedia Research 2019, doi.org/10.6084/m9.figshare.8214521 (CC BY 4.0)

¹ https://www.wikimedia.org/wiki/Strategy:Wikimedia_movement_2017/Direction
² <https://www.wikimedia.org/wiki/Strategy:Augmentation>
³ <https://www.wikimedia.org/wiki/Strategy:Augmentation>
⁴ <https://www.wikimedia.org/wiki/Strategy:Augmentation>

Ethical & human centered AI
whitepaper on [Figshare](#)

meta.wikimedia.org/wiki/Research:Ethical_AI

So far, I've taken an approach that is both top down and bottom up. This approach is described in detail in a white paper I published earlier this year called "Ethical & Human Centered AI". On the top-down side, I've identified a set of what I call proposals that cover some of the major pieces of actionable guidance for improving our product development process in order to support ethical outcomes, distilled from the relevant literature. On the bottom-up side, I've developed a set of six scenarios that describe unintended consequences which could potentially be avoided by adopting the proposals, each of which is anchored in a Wikimedia-specific AI product context.

Proposals

AI product development process

1. Checklists & impact assessments
2. Prototyping & user testing
3. Piloting & evaluation metrics

AI product design

4. Interpretable models & dataset documentation
5. UI explanations & user control
6. Auditing & feedback mechanisms



The proposals fall into two types: in the first category are recommendations for activities Wikimedia can perform during our design and development process, like developing ethics-focused checklists and piloting all products before committing to full-scale or long-term deployment. In the second category are design patterns or product features that we should adopt, like using interpretable models and providing product-specific mechanisms for auditing and feedback.

Scenarios

Short vignettes that describes the release of an AI product, and its impacts (positive and negative)

- Concrete
- Realistic
- Generative



The second technique I've been using to help identify ethical AI requirements for Wikimedia from the bottom up is developing scenarios, short vignettes that describe the release of an AI-driven product, and its impacts (both positive and negative). I've tried to design these scenarios to be concrete, realistic, and generative. Concrete because they describe specific products with a defined audience, purpose, and context; realistic in that the choice of product, its goals, and the unintended consequences highlighted in the scenario are all plausible given Wikimedia's goals and our prior knowledge about potential characteristics and impacts of related AI technologies. And generative so that they can serve as shared artifacts that facilitate discussion of risks, tradeoffs, assumptions, and remediations that are salient to people with different roles and backgrounds and are potentially relevant to other AI product contexts as well.

Scenarios

1. Reinforcing gender bias in content recommendation
2. Introducing cultural bias in reading recommendations
3. Community disruption through machine translation
4. Impacts of automated draft quality classification on diversity
5. Transparency and recourse in vandalism detection
6. Accountability for consequences of external re-use of labelled data



The six scenarios I describe in the white paper cover a range of potential unintended consequences related to different AI products Wikimedia develops, or that it might want to develop in future.

Scenarios

1. **Reinforcing gender bias in content recommendation**
2. Introducing cultural bias in reading recommendations
3. Community disruption through machine translation
4. Fairness and diversity impacts of quality classification
5. Transparency and recourse in vandalism detection
6. Accountability for consequences of external re-use

WIKIMEDIA
FOUNDATION

Today in the interest of time I'm going to focus on a single scenario.

Questions for each scenario

1. Could we have *anticipated this outcome* before we built and deployed the product?
2. Could we have *identified this outcome*, before or after deployment?
3. Could we have *achieved this goal* in a way that entirely avoided this possibility?
4. Can we think of *other AI products* where similar issues might arise?



As we go through the scenario, I encourage you to keep the following questions in mind:

1. How could Wikimedia have anticipated the negative outcome implicated in the scenario before we developed and deployed the product?
2. How could we have identified this negative outcome, before or after we deployed the product?
3. How could we have achieved the goal of the product while avoiding the negative outcome?
4. What are some other cases where this kind of outcome is a real risk?

Scenario 1

Reinforcing gender bias in Wikipedia biographies

First, a quick background on the scenario. Roughly 60-80% of articles on English Wikipedia are very short, and in many cases incomplete--meaning there's a substantial amount of relevant information about the article topic out there in the world somewhere which is not currently included in the article.

From Wikipedia, the free encyclopedia

Born in [Antwerp](#), Alice was one of three children. She was trained as a [dressmaker](#), a trade which she later taught as a school teacher. As with many Belgian families, the Freys left their home in the [First World War](#), moving to [Ostend](#). There Alice met painter [James Ensor](#) who became a close friend. Ensor encouraged Alice's interest in art, and after the war she enrolled as student in the [Academy of Fine Arts in Antwerp](#), where she studied drawing and painting.

Alice Frey's work was widely exhibited in her lifetime. In her later years she became profoundly deaf and lived alone in Ostend. On her death in 1981, much of her work was sold at auction, and only the small collection that she willed to Ostend is accessible. Stylistically, Frey's work was influenced by [Marc Chagall](#), Ensor, and [Edgard Tytgat](#). Her early works are [expressionist](#), while her later work may be closer to [magic realism](#).

• Dupont, Pierre-Paul, "FREY, Alice, épouse MARLIER" in E. Gubin, C. Jacques, V. Piette & J. Puissant (eds), *Dictionnaire des femmes belges: XIX^e et XX^e siècles*. Bruxelles: Éditions Racine, 2006. [ISBN 978-2-87386-434-7](#)



Born	June 25, 1895, Antwerp, Belgium
Died	August 30, 1981 (aged 86) Ostend, Belgium
Nationality	Belgian
Education	Academy of Fine Arts in Antwerp
Known for	Painting
Movement	Expressionism
Spouse(s)	Georges Marlier (m. 1922–1968)

Authority control 

BNF: [cb155967735 \(data\)](#) · [ISNI: 0000 0000 1679 7983](#) · [LCCN: no204086944](#) · [RKD: 29426](#) · [SNAC: w66x3vf9](#) · [SUDOC: 086028324](#) · [ULAN: 500139980](#) · [VIAF: 52751379](#) · [WorldCat Identities \(via VIAF\): 52751379](#)

This article about a *Belgian* painter is a *stub*. You can help Wikipedia by *expanding it*.

Here's an example of one of those so-called "stub" articles, this one about the Belgian artist Alice Frey. You can see that it's only a few paragraphs long, and contains very little information about the artists' work.

Alice Bailly

From Wikipedia, the free encyclopedia

Alice Bailly (25 February 1872 – 1 January 1938) was a radical [Swiss](#) painter, known for her interpretations on [cubism](#), [fauvism](#), her wool paintings, and her participation in the [Dada](#) movement. In 1906, Bailly had settled in [Paris](#) where she befriended [Juan Gris](#), [Francis Picabia](#), and [Marie Laurencin](#), avant-garde [modernist](#) painters who influenced her works and her later life.^[1]

Contents [hide]

- 1 Family and background
- 2 Education and early career
- 3 Inspiration and Fauvism
- 4 Cubism and wool paintings
- 5 Dadaism
- 6 *Salon de Independents*
- 7 Famous works
- 8 Later life
- 9 Retrospective Exhibitions
- 10 Notes
- 11 Further reading

Family and background [edit]

Originally, the family name was Bally, but after a critic mistook her name for "Bolly" in a review she had it changed to "Bailly" to avoid further confusions. She was born to a modestly situated family in [Geneva, Switzerland](#). Bailly's father, who worked as a Post Office official, died when Bailly was fourteen. Her mother, a German teacher, taught Bailly and her two sisters to be cultured and full of energy.^[2]

Education and early career [edit]

At seventeen, she attended the [École des Beaux-Arts](#) and took women's-only courses. She believed that the purpose of the school was to develop her individual

Alice Bailly



"Self Portrait" 1917

Born	25 February 1872 Geneva, Switzerland
Died	1 January 1938 (aged 65) Lausanne, Switzerland
Nationality	Swiss
Education	Oregeville Institute, Paris, 1910–11; École des Beaux-Arts, Geneva, 1891–95
Known for	Painting

Contrast this with the article about a roughly contemporaneous Swiss artist Alice Bailly. This article has lots of information, which you can see in the list of sections in this screenshot. Section headings like these serve as guideposts for content on Wikipedia articles. They not only provide a more navigable overall structure for the reader, they also reflect a kind of consensus about the kinds of information that is relevant for an article on a particular topic. So for example, more complete biographical articles like these will often contain a mix of sections related to both the personal and professional lives of the subject.

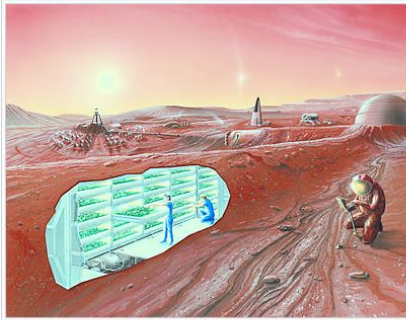
Another virtue of section headers is that they reflect somewhat regular patterns that can be leveraged by machine learning models. And in fact, Wikimedia has recently developed a model that predicts which sections a given stub article "should" have, based on the existing section headings and other textual characteristics of similar articles.

Colonization of Mars

From Wikipedia, the free encyclopedia

Mars is the focus of much scientific study about possible human colonization. Its surface conditions and the presence of water on Mars make it arguably the most hospitable of the planets in the Solar System, other than Earth. Mars requires less energy per unit mass (delta-v) to reach from Earth than any planet except Venus.

Permanent human habitation on a planetary body other than the Earth is one of science fiction's most prevalent themes. As technology has advanced, and concerns about the future of humanity on Earth have increased, the argument that space colonization is an achievable and worthwhile goal has gained momentum.^{[1][2]} Other reasons for colonizing space include economic interests, long-term scientific research best carried out by humans as opposed to robotic probes, and sheer curiosity.



An artist's conception of a human Mars base, with a cutaway revealing an interior horticultural area

Sections you can add

- [Relative similarity to Earth](#)
- [Differences from Earth](#)
- [Conditions for human habitation](#)
- [Radiation](#)
- [Transportation](#)
- [Equipment needed for colonization](#)
- [Robotic precursors](#)
- [Mission concepts](#)
- [Economics](#)
- [Possible locations for settlements](#)
- [Planetary protection](#)



[Edit to add new sections](#)

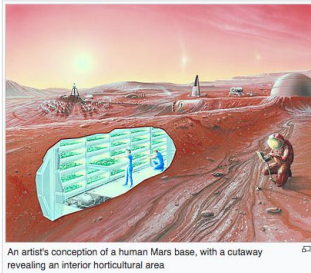
One practical use of a model like this, in an AI product context, is an interface that surfaces "recommended" sections to editors when they are editing a particular article. The screenshot above shows a conceptual mockup of what this kind of interface might look like.

Colonization of Mars

From Wikipedia, the free encyclopedia

Mars is the focus of much scientific study about possible human colonization. Its surface conditions and the presence of water on Mars make it arguably the most hospitable of the planets in the Solar System, other than Earth. Mars requires less energy per unit mass (delta-v) to reach from Earth than any planet except Venus.

Permanent human habitation on a planetary body other than the Earth is one of science fiction's most prevalent themes. As technology has advanced, and concerns about the future of humanity on Earth have increased, the argument that space colonization is an achievable and worthwhile goal has gained momentum.^{[1][2]} Other reasons for colonizing space include economic interests, long-term scientific research best carried out by humans as opposed to robotic probes, and sheer curiosity.



An artist's conception of a human Mars base, with a cutaway revealing an interior horticultural area.

Sections you can add

[Relative similarity to Earth](#)

[Differences from Earth](#)

[Conditions for human habitation](#)

[Radiation](#)

[Transportation](#)

[Equipment needed for colonization](#)

[Robotic precursors](#)

[Mission concepts](#)

[Economics](#)

[Possible locations for settlements](#)

[Planetary protection](#)

Goals

- Increase content quality
- help new editors learn-by-doing

Measures of success

- Overall growth in average article size
- Improved within-topic consistency
- increased new editor retention

WIKIMEDIA
FOUNDATION

From a product perspective, introducing a feature like section recommendations to Wikipedia could have several plausible goals: for example, a goal of increasing overall content quality across Wikipedia by encouraging people to expand incomplete articles; as well as improving the overall consistency in the structure and coverage of articles on the same topic. A feature like this could also potentially increase retention among new editors, a perennial problem for Wikipedia, by providing an engaging and educational contribution opportunity for new editors who might otherwise be unsure about how to add to articles that need improvement.

Scenario 1

Alice Frey


From Wikipedia, the free encyclopedia

Alice Frey (25 June 1895 – 30 August 1981) was a [Belgian painter](#).

Born in [Antwerp](#), Alice was one of three children. She was trained as a [dressmaker](#), a trade which she later taught as a school teacher. As with many Belgian families, the Freys left their home in the [First World War](#), moving to [Ostend](#). There Alice met painter [James Ensor](#) who became a close friend. Ensor encouraged Alice's interest in art, and after the war she enrolled as student in the [Academy of Fine Arts in Antwerp](#), where she studied drawing and painting.

It was as a student that she met her future husband [Georges Marlier](#), himself later a noted art critic and painter. They married in 1922. Together they formed part of a group known as *Lumière*, which published a journal. A second journal, *Ça Ira*, was established by Alice, and these formed part of the avant-garde movement in Belgium in the 1920s.

Alice Frey's work was widely exhibited in her lifetime. In her later years she became profoundly deaf and lived alone in Ostend. On her death in 1981, much of her work was sold at auction, and only the small collection that she willed to Ostend is accessible. Stylistically, Frey's work was influenced by [Marc Chagall](#), Ensor, and [Edgard Tytgat](#). Her early works are [expressionist](#), while her later work may be closer to [magic realism](#).



Alice Frey

Born June 25, 1895, Antwerp, Belgium

Died August 30, 1981 (aged 86) Ostend, Belgium

Nationality Belgian

Education Academy of Fine Arts in Antwerp

Known for [Painting](#)

Movement [Expressionism](#)

Spouse(s) [Georges Marlier](#) (m. 1922–1968)

Sections you can add

- [Personal life](#)
- [Influences](#)
- [Family](#)
- [Later life](#)
- [Early career](#)
- [Education](#)
- [Artistic style](#)

[Edit to add new sections](#)

Wikimedia builds a section recommendation widget to help editors expand stub articles.

The underlying recommender model learns that biographies of women tend to have sections with titles like “Personal life” and “Family”, while biographies of men have sections like “Career” and “Awards and honors”.

It makes section recommendations based on what it has learned.

So imagine that Wikimedia builds this AI product, and it becomes popular among new editors and is widely used. The section recommendation model, trained on the corpus of sections in existing biographical articles, learns among other things that biographies of women tend to include sections with titles like "Personal life" and "Family", while biographies of men are more likely to contain sections like "Career" and "Awards and Honors".

If this feature is widely adopted, it is likely to achieve its stated goals. However, it is also likely to result in the unintended consequence of reinforcing existing gender biases in the way women are represented on Wikipedia. And over time, the more the feature is used, the more biased Wikipedia becomes with respect to gender.

Unpacking scenario 1: Gender bias in Wikipedia

“articles about women tend to emphasize the fact that they are about a women (i.e., they contain words like ‘woman’, ‘female’ or ‘lady’), while articles about men don’t contain words like ‘man’, ‘masculine’ or ‘gentleman’.

“words like ‘married’, ‘divorced’, ‘children’ or ‘family’ are much more frequently used in articles about women... in the English Wikipedia **an article about a notable person that mentions that the person is divorced is 4 times more likely to be about a woman** rather than a man.”

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science*, 5(1), 5.

Here I should note that with respect to gender bias in biographies, we could probably predict that something like this could happen. The gender bias in biographies has been noted before in other features of articles, so it’s reasonable to expect it might show up in section headings.

Before I close out the scenario I'd like to highlight a few considerations that make it more difficult, from a product perspective, to address issues like this one than it might initially seem.

Unpacking scenario 1: Challenges

- **Impact evaluation:** time scales and effect sizes
- **Retrospective remediation:** preserving value added and avoiding wasted work
- **Prospective remediation:** monitoring the outcomes of de-biasing efforts
- **Transparency and trust:** do we have a duty to report?
- **Unknown unknowns:** what *other* biases are we reinforcing?



One issue is related to evaluation. If the overall pace of new content development is slow enough, and the bias effect subtle enough, that it could take a long time to demonstrate an effect. Perhaps far longer than even a data-driven technology company is willing to keep a product in beta, pilot, or small-scale deployment.

Another issue is related to remediation. If the bias is uncovered long after the "damage" has been done, how do we undue the bias without also throwing out the valuable work performed by volunteer editors who are invested in the content they've created?

How should we go about de-biasing the model? And how do we know that the de-biasing is effective at addressing the problem, or whether that effort has caused additional unintended consequences of its own?

Finally, how do we identify unknown unknowns? In this particular case, we may be able to draw on previous research to develop methods for characterizing and quantifying the negative impact--assuming someone in the organization is aware of the literature and they have a voice in design discussions. But even if we are aware of the risk of reinforcing bias in a this particular subset of articles and have an idea of what to look for, how do we account for sources of bias that no one has even bothered to write a research paper about?

English Wikipedia happens to have many such biases, due to the composition of the community of editors that write articles, which is primarily male, highly educated, and North American and Western European. Biases have also been identified in terms of how topics relevant to different cultures and geographies are presented (as well as what gets covered).

We are aware of some of these biases, but we know there are likely many more we aren't aware of, or that we don't know how to measure.

Unpacking scenario 1: Trade-offs

- **Prioritization:** which potential sources of bias are most important?
- **Benefit vs. harm:** when does a global improvement outweigh a local regression?



The range of potential bias issues, and the risk of inadvertently making them worse with machine learning, highlights some important trade-offs. If we suspect that bias is endemic do we prioritize what potential sources of bias address first, or at all? We can expect that section headings will embed many other kinds of bias across different topics that we would view as problematic if we knew about them, but how do we prioritize which ones to focus on, and how do we know what to look for as a signal that bias is being propagated through use of this AI product?

Furthermore, how do we weigh the risks of increasing bias in some locales, with the opportunities that machine learning presents to improve the quality of articles across Wikipedia--even, potentially, articles where it also introduces bias? How do we weigh the risk of creating more bias in some sub-sets of articles, with the potential advantages of increasing new editor retention? ~~FIXME~~ How do we quantify the relevant advantages of this AI product on overall article growth and quality--including on biographical articles about women, which are disproportionately stubs after all--with the risk of perpetuating harmful societal stereotypes?

Product development priorities & tensions

- **Product planning:** Opportunities missed vs. bullets dodged
- **Product development process:** up-front costs vs. lifecycle costs
- **Product evaluation:** Direct impacts vs. second-order effects
- **Organizational change:** Switching costs vs. continuous improvement



The theme of trade-offs takes us to the final part of my talk today. The choices that Wikimedia, or any technology organization, makes when faced with trade-offs like these at the level of an individual product reflects important considerations about its product development process, and the nature of the organization itself. Such as whether the organization chooses to optimize for avoid missed opportunities vs. avoiding unintended negative consequences; whether the organization is willing to incur tangible costs in terms of time spent planning, iterating, and testing products up front in order to avoid the potential cost of fixing or withdrawing problematic products down the line. Whether the organization chooses to prioritize measuring performance against direct and intended outcomes vs. unintended consequences and other second-order effects; and how it balances the value of opportunities for improvement that result from experimenting with new approaches to design and development with the switching costs that result from disrupting existing software development workflows.

Guiding questions redux

1. What steps can Wikimedia realistically take now, with our current resources?
2. What initial steps position Wikimedia to iterate, improve, and evolve?
3. Is product development at Wikimedia fundamentally *goal*-driven or *values*-driven?



In balancing these tradeoffs and prioritization decisions with the original guiding question for this work "what would an MVP for ethical AI at Wikimedia look like?" I've come to three new guiding questions, which I've begun to use, along with the scenarios and process proposals, to begin discussions with our research scientists and product teams:

1. What process improvements can we implement now, given the personnel, priorities, products, and expertise we currently have?
2. Which initial improvements are most likely to be generative in themselves, and will help us continue to learn as we expand and scale our ethical AI efforts?
3. And finally, fundamentally, should our product development be more goal-driven--meaning we achieve the things we set of to achieve, or values-driven--meaning that the outcomes of our decisions align with and perpetuate the values that we claim to hold?



Thank you!

Jonathan T. Morgan

Workshop on AI at Wikimedia • Berkman Klein Center • 25 June 2019

meta.wikimedia.org/wiki/Research:Ethical_AI

Project page:

https://meta.wikimedia.org/wiki/Research:Ethical_and_human-centered_AI