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 wellbeing throughout product development.

This white paper is intended to help Wikimedia ensure ethical and human-centered outcomes in AI product development given our current and anticipated goals, needs, capacities, and workflows. The paper makes two contributions: 1) it motivates a set of *risk scenarios* intended to define the problem space and promote reflective decision-making, and 2) it presents 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 2017 Strategic Direction.[1] It complements the strategic priorities described in the white papers *Knowledge gaps, Knowledge integrity,* and *Foundations*[2] by Wikimedia Research and *Augmentation*[3] by Wikimedia Audiences. Background material and additional resources are available on meta.wikimedia.org.[4]

Cite this document as: Jonathan T. Morgan, 2019. *Ethical & Human Centered AI - Wikimedia Research 2030*. doi.org/10.6084/m9.figshare.8044553 [CC BY 4.0]

---

[1] https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017/Direction
[2] https://meta.wikimedia.org/wiki/Research:2030
[3] https://www.mediawiki.org/wiki/Wikimedia_Audiences/Perspectives/Augmentation
[4] https://meta.wikimedia.org/wiki/Research:Ethical_and_human-centered_AI

[Version 1.0 • April 25, 2019]

Ethical AI: opportunities and challenges

The Wikimedia Foundation already develops products that utilize machine learning, or *AI*, technology in some important contexts: to detect vandalism, improve search, assess articles, and make translation recommendations.

But this is only the beginning. Products powered by AI technology have the potential to address many of the other challenges implicated in Wikimedia's 2030 Strategic Directions of *Knowledge Equity* and *Knowledge as a Service*. Imagine a world in which machine learning augmented workflows help contributors find useful tasks that match their interests and their expertise, whether they're a newbie or a veteran; in which computer vision tools help readers and editors find, sort, and annotate images and media; and in which WMF product teams and movement volunteers are able to dynamically visualize clusters and hierarchies among texts, media, and structured data entities—helping them identify gaps that need to be filled in.

However, the scale, speed, and complexity of AI technology present risks and challenges beyond those we typically encounter when we build non-AI software. Automating or augmenting established workflows with machine learning can disrupt, distract, and disenfranchise people; facial recognition technologies may misidentify people with darker skin or even fail to recognize them as human; algorithmic recommendation and classification systems can create 'filter bubbles' and often rely on collecting and storing vast amounts of potentially sensitive usage data and personally identifiable information.

If Wikimedia wants to reap the benefits of AI technologies without damaging our projects, communities, mission, or reputation, we need to adapt our processes for planning, designing, deploying, and evaluating new software products.

Defining and designing AI products

At Wikimedia, we don't build standalone AI products—we build *AI product ecologies*. **Table 1** presents a non-exhaustive list of distinct, but interdependent, AI products that were developed to help editors harness machine learning to identify and deal with vandalism. Some of the products are used to develop and maintain machine learning models, others help people use the output of those models. Each was designed with particular users, needs, and use cases in mind. Each is a product in its own right—regardless of whether it has a graphical user interface, or was built by a traditional product team.

A lot of product development at Wikimedia happens outside of traditional product teams. We actively encourages our volunteer community to build new products on top of those we design ourselves—it's part of our ethos and our shared movement governance model, and it's also a practical way to scale our work. Our products—user interfaces, APIs, datasets, and machine learning models—are available for anyone to use. Sometimes, people use these products for purposes we never anticipated—for example, to train a *Jeopardy*-winning AI like IBM's Watson.

Openness allows our movement to innovate at scale and respond agily to emergent needs, but it can also lead to unintended consequences. The risk of unintended consequences is high with AI products, where small changes in one product's code, training data, documentation, or user interface can have unpredictable impacts on the way downstream products function, or on how people use them. Many of the negative impacts of AI products can be traced upstream to previous decisions made about *other* AI products—decisions made by a different person (or team), often for a different reason.

Wiki Labels	An interface for crowdsourcing data labelling tasks
Labeled data	A dataset of labelled revisions
Revision Scoring	A machine learning model trained on a labelled dataset
ORES	An online platform that hosts and provides API access to model predictions
New Filters for Edit Review	A configurable edit feed interface augmented with model predictions
JADE	An interface for disputing model predictions and re-labeling data

Table 1. AI products related to edit classification

Entry points for bias and harm

In order to build AI technologies in an ethical and human-centered way, we must consider the full stack of AI products we create around those technologies. Inherent biases in training data, unclear documentation, black box machine learning models, and a product designer's own assumptions about what is true and what is important can all lead to bad outcomes.

Cases of algorithmic bias often arise from bias in the training data. Machine learning models trained on data created by humans reflect the assumptions, understandings, and biases of those humans. A machine learning model trained on a dataset of accepted and rejected resumes that contain demographic information (e.g. gender, age) will likely reflect the level of sexism and ageism that motivated the managers who made the hiring decisions—whether or not the managers or model builders were aware of the bias. A resume screening software product based on such a model could easily be unfairly biased against women and older job candidates. How would an HR representative using that product become aware of the risk?

Upstream design decisions can impact downstream products in profound ways. When AI products are built using a deep learning approach it becomes much more difficult to predict or explain why the model arrived at a particular decision. If a software program designed to help doctors identify patients who are in imminent danger of stroke is unable to flag the relevant risk vectors for them, how can that doctor determine the best course of treatment?

Re-use of AI products outside the original context can have unpredictable results. A machine learning model designed to predict whether a sentenced criminal is likely to reoffend can lead to unfair outcomes (and potentially legal trouble) when used to decide who to give insurance to. Wikimedia doesn't control how people re-use the data and models that we release. What ethical responsibility do we do bear for the consequences? How do we assess the risk that re-use will conflict with our values, undermine our goals, or result in harm to people inside or outside our movement? How can we help people make informed decisions when they re-use our products in "off-label" ways?

Risk scenarios

Ultimately, we cannot hope to account for every use case, edge case, and dependency within our AI product ecosystem. We cannot predict with 100% accuracy what negative impacts the design of an API, model, user interface, or dataset documentation will have downstream.

However, we can anticipate, prevent, and mitigate many negative consequences by following best practices for ethical and human-centered AI product design. In order to understand what design process changes are necessary, we start by outlining a set of plausible scenarios that describe unintended consequences of well-intentioned AI products that are similar to the ones we currently have.

The scenarios below are *fictional*. They are designed around realistic AI products and product *use cases* within the Wikimedia movement, but they are not statements of fact or findings from empirical research. They are intended to illustrate some of the bad outcomes that might result from seemingly sensible design decisions made by different people at different points in the product development process.

These scenarios are meant to show how good faith attempts by intelligent people to leverage machine learning to improve Wikipedia can still result in outcomes that run counter to our mission and our values.

I assert that preventing outcomes like these is the responsibility of the organization as a whole, acting together—not of any particular product manager, designer, researcher, or engineer. How the organization should exercise that responsibility is the question that the rest of the paper attempts to answer.

When reading the scenarios below, consider the following questions:

- **Identifying risks:** How could someone involved in the development of this product have identified this risk beforehand?
- **Preventing outcomes:** What could they have done to avoid this outcome before release?
- **Mitigating impacts:** How could they have detected and addressed this bad outcome after release?
- **Assessing responsibility:** To what extent, and in what way, should Wikimedia hold itself responsible for anticipating and addressing the consequences of this scenario?

Scenario A. Reinforcing existing bias

Wikimedia builds a section recommendation feature into the editing interface. This feature uses machine learning to suggest a list of potential section headings for very short articles—based on the sections that already exist in other articles that resemble them—creating hooks to encourage article expansion.

The section recommender learns that Wikipedia biographies about men are likely to have section titles like "Career" and "Awards and honors", while biographies of women are more likely to have sections with titles like “Personal life” and “Family”. The feature is widely used: it increases the overall quality of short Wikipedia articles, but also increases the systemic bias in the way women and men are portrayed on Wikipedia.

Scenario B. Discouraging diversity

Wikimedia adds a draft quality score into the new article review dashboard. The quality prediction model weighs spelling errors and grammatical disfluencies highly when scoring articles, but doesn’t consider the number of citations or the notability of the topic. As a result, articles written by people for whom English is a second language tend to have lower overall scores.

Many content gaps in Wikipedia are around topics that native English speakers tend to be less knowledgeable about or interested in. However, reviewers find the scores work well enough and they can make acceptance and rejection decisions much faster based on the score alone. As a result, reviewer workload is decreased, but good quality new articles on important topics are rejected at an increased rate, and culturally diverse contributors frequently see their hard work deleted.

Scenario C. Lack of transparency and recourse

Wikimedia builds filters powered by machine learning into the recent changes, article history, and watchlist feeds on Wikipedia. One of these filters highlights edits that have a high probability of being performed with malicious intent.

A 5-year veteran editor with over 50,000 edits notices that many of their recent edits have been highlighted as likely bad faith, and that their edits are now being reverted at a much higher rate than before the filters were rolled out. After some sleuthing, they notice some patterns and believe they have figured out why their edits are being erroneously tagged as the result of a ‘corner case’—a rare combination of factors related to the kinds of edits they make and the articles they work on is confusing the model. However, they are unable to discover a way to report the issue, correct the faulty predictions, or confirm their suspicions about the cause. In the meantime, the heightened level of scrutiny and rejection they experience from their fellow editors leads to embarrassment, conflict, a sense of alienation, and they consider leaving Wikipedia.

Scenario D. External re-use and harm

Wikimedia releases a dataset of Wikipedia talk page comments, labelled by crowdworkers for key words and phrases related to toxic speech. An external developer trains a machine learning model on this dataset, and uses to model to power an automated content moderation system for an online depression support forum for at-risk teens.

Teens experiencing mental health crises tend to use emotionally charged language and words commonly associated (in other contexts) with aggression and hate speech. While the automated system proves effective at deleting trolling posts quickly, it also flags and deletes many legitimate support requests that are appropriate to the forum and permitted under its rules. The messages that demonstrate the greatest need for support are the most likely to be blocked by the tool.

Scenario E. Community disruption and cultural imperialism

Wikimedia builds a tool that recommends articles to translate from one language to another. The tool uses machine translation to generate pretty-good translations of articles. The goal is to leverage the content and community of large Wikipedias to help smaller Wikipedias grow.

The translations are good enough that editors from English Wikipedia who know a little bit of Amharic feel confident using this tool to publish lots of articles to Amharic Wikipedia after a quick review and some light clean-up. They appreciate the opportunity to find valuable work that matches their interests and their expertise. Amharic Wikipedia is much smaller than English, has fewer editors overall, and fewer bilingual editors, and the local community is currently focusing on expanding their Wikipedia organically and curating the content they already have.

The Amharic community soon finds themselves overwhelmed by an influx of new imperfectly-translated articles. Although the Encyclopedia grows faster, local editors must now focus their energy on fixing errors and completing partial translations, rather than writing the articles that they are interested in writing, or that they believe are most important to their readers.

Scenario F. Cultural assumptions and disparate usefulness

Wikimedia deploys a new ranking algorithm to power the *top articles* feed in the English Wikipedia Android App. The previous ranking algorithm was based on a simple pageview-based metric: it reflected what Wikipedia readers are interested in reading. The new ranking is based on a more sophisticated machine learning model that predicts which articles are 'trending' based on patterns of editing activity associated with breaking news events: it reflects what Wikipedia editors are interested in editing.

Most English Wikipedia editors are North American or Western European. Many members of the app development team are English Wikipedia editors. For them, the new ranking seems to perform better: it surfaces trending articles that are more relevant to their interests.

However, a large proportion of mobile English readers come from India. These Indian readers value the pageview-based feed because it frequently surfaces articles that are culturally relevant to them. After the new algorithm is deployed, the feed contains fewer articles these readers find interesting. Over time, they unconsciously begin using the Wikipedia app less frequently in favor of information sources that reflect their interests and meet their needs better.

Unpacking the scenarios

None of the hypothetical bad outcomes described above were inevitable. Previous research has demonstrated that notable women and men are written about in distinctly gendered ways (**Scenario A**). Draft quality review scores (**Scenario B**) could have been released on a trial basis, with the final adoption of the technology made contingent on an impact assessment on new editor retention. Clear reporting channels, or community-run dispute resolution or emergency shutoff processes would have better supported the editor accidentally targeted in **Scenario C**. More detailed documentation on data provenance and intended use cases could have helped the external developers who reused the toxicity dataset (**Scenario D**) better understand its limitations, and design their AI product accordingly. User research—interviews, surveys, or lab studies—would likely have uncovered evidence that an influx of poorly translated and culturally irrelevant articles (**Scenario E**) could be disruptive for a smaller Wikipedia community. Comparative, bias-focused algorithmic audit studies using real data and crowdworkers recruited from different geographies could have identified culturally-mediated differences in the perceived relevance and usefulness of recommended content (**Scenario F**).

All of the hypothetical unintended outcomes in the scenarios above point to outcomes that undermine Wikimedia's strategic goals. However, that doesn't mean that all of these outcomes are equally *bad*. Some of these scenarios may reflect acceptable risks or necessary trade-offs. But making trade-offs implies awareness of potential costs and benefits beforehand. How can the Wikimedia Foundation, as an organization, make more informed decisions around AI product development?

Human centered and ethical AI

This paper draws on human centered design and ethical AI to demonstrate how Wikimedia can build more innovative and impactful AI products while avoiding the kinds of unintended consequences described in the scenarios above.

Human centered design process

Human centered design consists of a set of research and design techniques that provide insights into user experience and inform product decision-making. Human centered design prioritizes contextual usability, utility, sustainability, accessibility, and human well-being. Human centered design research methods range from ethnographies and surveys to controlled experiments. Design methods range from brainstorming, persona development, and participatory design activities to iterative cycles of prototyping and testing.

A human centered design process is critical when building AI products because it creates opportunities to discover and correct issues throughout the product lifecycle, not just immediately before or after the product has shipped. The interventions proposed to identify, address, and mitigate unintended consequences in "Unpacking the Scenarios" above are all common approaches within human centered design.

Ethical AI principles and methods

Three basic principles of ethical AI are *fairness, accountability,* and *transparency*. Addressing systematic harmful bias is core concern for ethical AI that touches all of these principles. In this context, fairness means the absence of systematic bias, transparency means the ability to assess causes and consequences of bias, and accountability means taking responsibility for bias identification, prevention, and mitigation (i.e. providing redress for harm done, correcting errors to prevent future harm).

Ethical AI also focuses on other types of harm beyond bias, such as misuse of personal data, deception and manipulation, automating people out of a job, arbitrarily restricting users' choices, or people the information necessary to make informed decisions—including whether or not to use a product in the first place.

Process proposals

A human centered design process supports ethical AI principles by helping AI product designers think critically about success criteria, contingencies, and unintended consequences, and by maintaining alignment between design decisions and guiding principles throughout the product development lifecycle—before and during development, and, after deployment.

The eight proposals below were discovered through an extensive literature review and selected as candidate recommendations because they:

- *Are recommended by ethical AI scholars and scientists*
- *Align with Wikimedia's values, and*
- *Build off of established, effective practices already in place at Wikimedia*

These proposals are intentionally defined at a fairly high level—this document doesn't specify what should go on a checklist, or how many times you need to prototype, or what counts as a "meaningful" explanation of model behavior for a particular AI product. More details about these methods, potential pros and cons of each, and links to related work are provided on the wiki.

1. Use checklists during development

Related scenarios: B, C, F

An ethical AI checklist consists of a list of important steps that must be taken, or questions that must be answered, at each stage of the product development. Checklists work best when the process of working through the checklist is performed consistently, transparently, and collaboratively among team members. Used in this way, checklists create opportunities for open discussion among about assumptions, priorities, trade-offs, and risks. Checklists emphasize accountability and can help redistribute decision-making power within the team: if an important step or consideration is explicitly documented in the checklist, stakeholders with less power many be more comfortable flagging missed steps or raising other process issues without fear of reprisal.

2. Publish impact statements

Related scenarios: A, D, E

An ethical AI impact statement is a product plan that is published before substantial development begins. Impact statements include a detailed product rationale, supporting research, risk assessment, success criteria, and maintenance and monitoring plans. Impact statements increase accountability throughout the product lifecycle: accountability to address feedback from external stakeholders before development starts; to articulate assumptions and assess risks (such as the potential for bias) before launch; to evaluate success based on human-centered criteria; and to monitor for and address unintended consequences on an ongoing basis.

3. Prototype and user test

Related scenarios: C, F

Prototyping is a process of making iterative, incremental refinements to a design based on explicit feedback or observations of use before full deployment. Prototypes are usually lower fidelity than the final product: e.g. sketches, mock-ups, or simplified versions. Iterative cycles of prototyping and testing allows teams to verify their design ideas and their assumptions about product user characteristics, needs, and context of use before they commit to design decisions that would be costly or difficult to change later in the process. In software development, prototyping and testing is most commonly performed on software with graphical user interfaces—but dataset documentation, APIs, voice interfaces, and machine learning models can also be prototyped and tested with users.

4. Pilot and assess

Related scenarios: A, B, E

A pilot is a fixed-term or limited scale deployment of a final product, where the decision to fully deploy is deferred until the outcome of the pilot is assessed. Unlike prototypes, pilots involve putting finished products in front of real users and tracking how the product performs in the wild over an extended period of time. Pilot deployments require pre-specified success criteria, and a commitment from the team to revise or withdraw if those success criteria are not met, or if the product results in unanticipated consequences that cause harm to end users or other stakeholders.

5. Build interpretable models

Related scenarios: B, C, D

Interpretable machine learning models are models that a) expose the logic behind a particular output or decision, and/or b) expose the general features, procedures, or probabilities implicated in their decision-making in a way that the intended audience can understand. Models that are not interpretable, or *opaque*—the way the model makes a decision cannot be accounted for or reverse-engineered after the fact—are much more difficult to assess for harmful bias. Interpretable models allow model designers, external auditors, and (via UI explanations, below) end users to evaluate whether a model is meeting expectations and requirements.

6. Provide UI explanations

Related scenarios: B, C, D

UI explanations consist of contextual metadata about how a model works that is made available to end users at the point of use. UI explanations can be written in words, or presented as statistical probabilities or graphical visualizations. The primary purpose of a UI explanation is to provide an end user with enough information to meaningfully understand how the model works in general, and/or how a particular model decision was made, so that they can make informed decisions about how much to trust the output of a machine learning model, or whether/when to use it.

7. Support external auditing

Related scenarios: A, B, E

Auditing is the process by which entities outside of the AI product team also inspect and spot-check the inputs and outputs of a machine learning model, or critically evaluate the design process behind an AI product. From an ethical AI perspective, the point of auditing is ensuring fairness: auditors may look for systemic biases in the way a model performs or identify conditions where a software tool built around a model may exhibit bias against particular individuals or groups. Auditing support can take the form of providing open APIs that allow arbitrary selection of inputs, feature weights, or other conditions and examination of the raw output. Building interpretable models and making training datasets and detailed process documentation available also supports auditing.

8. Provide built-in feedback mechanisms

Related scenarios: C, E, F

Feedback mechanisms allow product users to correct, contest, refine, discuss, or dismiss the output of a machine learning model at the point of use. Feedback can be qualitative (e.g. free text comments like bug reports) or quantitative (e.g. relevance ratings for “articles you might be interested in” recommendations). Feedback can be consumed and triaged by product team members, tracked automatically over time like a KPI, or incorporated directly into a model’s training data—retraining the model to refine it predictions or customize them for a particular user need.

Feedback mechanisms support continuous iterative improvement after deployment. Edge and corner cases where an AI product fails to meet user needs, or exhibits harmful bias, may not be frequent or obvious enough to notice through user testing or a small-scale pilot deployment. Machine learning models are susceptible to *drift*—a model that was accurate when launched can become less accurate or useful over time due to changes in the environment, the user base, or the underlying technology stack.

Unpacking the process proposals

The proposals above provide targeted guidance on integrating ethical and human centered AI into all stages of the product lifecycle: from brainstorming and project planning through design, deployment, and maintenance. The impetus for developing these proposals—and the white paper as a whole—was the question *what would a minimum viable ethical AI product development process at Wikimedia look like?* As such, these aren’t intended to be the final word on what WMF should, or shouldn’t, do when deciding what products to build or how to build and evaluate them.

Organizational implications

If Wikimedia decides to implement some or all of these proposals, what kind of impacts can we expect? The proposals were selected in part because they build off of existing organization strengths, such as:

1) Our *commitment to openness*—we already document much of our work in public, and publicly share much of the code and data that power our products;

2) The *expertise of our movement volunteers* who provide valuable feedback and often serve as co-designers of our products;

3) Our *mission-driven metrics* which lead us to define success by increasing the reach and quality of Wikimedia projects the and health and diversity of our contributing communities, not by profit or market share.

4) The high *trust in the Wikipedia brand* that we enjoy among readers.

Taken together, these strengths endow us with a great deal of freedom to experiment with new approaches to AI product design, and perhaps a greater degree of *permission to fail* than many other technology companies—as long as we can justify our product decisions in the terms of our mission and values. Taking an explicitly human centered and ethical approach to development will make it much easier for us to make this kind of justification.

Key challenges

Implementing these process proposals in a rigorous, systematic, and sustainable way will require organizational change, and organizational change is always a challenge.

If these proposals are adopted, one anticipated challenge for AI product development will be adjusting to longer product development timelines. The proposals above are intended to standardize product development and introduce new phases in the product lifecycle. Front-loading research and risk assessment in product development, requiring iterative cycles of prototyping and testing, committing to piloting before production—committing to performing these activities rigorously, systematically, and consistently is likely to create new dependencies across roles and teams.

On the other hand, if we follow the process proposals described above, we will probably find that the products we ship are more impactful and sustainable, and less likely to require sunsetting or extensive retrofitting after release—making our AI product development pipeline is more efficient, even if each individual products take longer to build.

Many of the successful AI products we have now are the result of individual initiative and serendipitous collaborations—we lack essential mechanisms for organizational learning, quality control, and scaling our development process. The development of AI products at Wikimedia takes place across multiple departments and teams. We currently lack dedicated capacities for coordinating the development of machine learning models with the development of products built around those models—especially when those activities occur in different departments. We also lack formal procedures for assessing the strategic fit, technological feasibility, or risk/reward matrix of proposed AI products.

Conclusion

Ultimately, the strongest argument for a more structured, iterative, and deliberate development process is that the ethos of “move fast and break things” clearly doesn’t work well for AI products. High-profile failures by other technology companies have resulted in a loss of public trust and concrete harm to individuals and society as a whole. In order to avoid eroding public trust, alienating our volunteer contributors, and betraying our core values, we must take concrete steps to ensure that the AI products we develop are ethical and human centered.

Engaging with risks scenarios like the ones provided during product development can help us—as individuals and as an organization—learn how to think about the ethical underpinnings and implications of our work. Implementing new AI product research, design, and development methods can help us achieve our strategic goals of increasing the reach and quality of Wikipedia and its sibling projects, and promoting diversity and growth within our volunteer contributor communities—while avoiding unintended consequences that conflict with our values, damage our reputation, and prevent us from building free cultural resources that everyone in the world can contribute to and benefit from.

Acknowledgments

This white paper was developed and refined through discussion and feedback from many people, including Jan Gerlach, Aaron Halfaker, Isaac Johnson, Jon Katz, Jess Klein, Margeigh Novotny, Miriam Redi, Diego Saez-Trumper, and Sherwin Sly.