Collective Meaning Cascades but Strange Ducks Swim Upstream

Facilitating Collective Meaning-making through Co-development of AI Models

Aaron L Halfaker
MSAI
Microsoft
Redmond, Washington, USA
aaron.halfaker@gmail.com

Tzu-Sheng Kuo
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
tzushenk@cs.cmu.edu

Ciell Brusse
unaffiliated
Amsterdam, Netherlands
ciell.wikipedia@gmail.com

Kenneth Holstein*
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
kjholste@andrew.cmu.edu

Haiyi Zhu*
Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
haiyiz@cs.cmu.edu

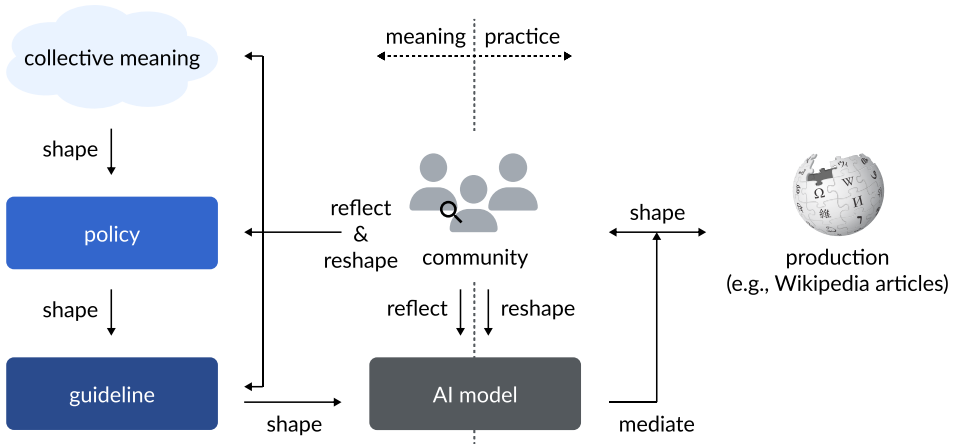

Figure 1: The collective meaning cycle. The meaning cascades through collective "meaning-making" processes into the formal document genres (policies and guidelines) and into AI models. In turn, each document and model is shaped by the flow of meaning and operate as mediators to practice. Meaning cascades through policies and guidelines and is re-shaped/translated/transformed in the process. Both the community and AI model straddle the line between *meaning* and *practice* and thus in the cycle provide a conduit for reflective processes to reshape the entire cascade. See Figure 5 for a visualization of the reflect/reshape process enabled by community co-development of models.

Abstract

Communities of practice operate by developing, sharing, and formalizing concepts together – collective meaning-making – thereby enabling all their community members to work together effectively. In the context of Wikipedia, these concepts include article quality, vandalism, and other subjective aspects of collective work. AI and machine learning have proven to be powerful tools for facilitating collaboration at scale by modeling and applying shared concepts. We examine meaning-making in parallel with aligning AI behavior. We describe a case study of modeling the quality of articles in Dutch Wikipedia using an AI model, while engaging in a meaning-making process with Dutch Wikipedians. This case blurs the line between social governance and how meaning is reshaped in an AI model. Based on the case study, we present the Collective Meaning Cycle, a framework that describes the bidirectional relationship between modeling and meaning-making. We also provide implications for the practice of participatory AI design.

*Co-senior authors contributed equally.

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3706683

CCS Concepts

• **Human-centered computing → Wikis**; **HCI theory, concepts and models**; • **Computing methodologies → Machine learning approaches**.

Keywords

Social Computing; Wikipedia; Machine Learning; Participatory Design; Ostrom; Governance; Meaning-making; Genre Ecology; Policy; Artificial Intelligence; Machine Learning

ACM Reference Format:

Aaron L Halfaker, Tzu-Sheng Kuo, Ciell Brusse, Kenneth Holstein, and Haiyi Zhu. 2025. Collective Meaning Cascades but Strange Ducks Swim Upstream: Facilitating Collective Meaning-making through Co-development of AI Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3706599.3706683

1 Introduction

The development and application of shared concepts is central to the functioning of communities of practice like Wikipedia. From Wikipedia's central pillars (e.g., Verifiability[1]) to shared understandings of what types of articles are *encyclopedic*, Wikipedians manage collaborative work by discussing, debating, recording, formalizing, and citing shared concepts in the form of essays, guidelines, and policies. This process of developing and capturing shared understandings is well described by the research literature [6, 20].

As Wikipedia scales, artificial intelligence (AI) and machine learning (ML) technologies have become core infrastructure for managing massive collaboration on the site. AI models are used to detect vandalism [1, 14], measure the quality of articles [25], route newly created articles to reviewers by topic [2], detect policy and style[2] issues in text [3], and even to generate encyclopedia articles directly from source material [16]. Without exception, each of these models is designed to "align" an algorithm's behavior with a tangible shared concept extensively documented in Wikipedia. The goal of AI alignment work is to steer AI models' behaviors toward these group norms [7].

In this paper, we report on a case study of building an AI model in the context where no documented norm was available to align the AI model. This case study allows us to unpack the relationship between meaning captured in norms. Through this unpacking, we contribute the *collective meaning cycle*, a framework that describes the bidirectional relationship between AI modeling and collective *meaning-making* (cf. [20]) within communities. The framework provides a deeper understanding of what it means to align algorithmic behavior in a social context. It provides implications for how AI developers might consider the design of well-aligned algorithmic systems in social contexts and adds a new thread to the conversations about genre ecologies [22] and articulation work [24] within open collaboration communities.

In the rest of the paper, we first review relevant literature and introduce our study method. Next, we describe the context of Dutch Wikipedia and their challenge of defining the article quality scale to model. We then highlight novel themes and insights that emerged throughout the model/meaning co-development process. Finally, we present the Collective Meaning Cycle and discuss its implications for AI alignment.

[1]https://enwp.org/WP:VERIFY

[2]See https://enwp.org/WP:VANDAL, https://enwp.org/WP:ASSESS, https://enwp.org/WP:NPOV, https://enwp.org/WP:WPDIR, and https://enwp.org/WP:MOS respectively

2 Related Work

2.1 Collective Meaning and Mediating Documents

In this paper, when we refer to *collective meaning*, we draw a connection to past work discussing the *collective meaning-making* [20] done in Wikipedia where community members engage in *articulation work* to build "shared understandings" of how to build and maintain an encyclopedia [6, 24]. Collective meaning represents the shared understanding about collective practice un-transformed, and therefore not distorted [15], by the act of translation into a formal document[3].

In order to be more easily shared and re-used, Wikipedians have created a formalized document genre [17, 22] called *policies and guidelines* as a foundational component of the distributed governance structure in Wikipedia [6]. These documents play a *mediating* role [17] by drawing connections between the *practice* of editing Wikipedia and *collective meaning*. Despite their drawbacks (inherent translation/distortion), these policy and guideline documents are useful as cite-able *mediators* of collective meaning [4].

Taken together, the literature paints a clear picture of how *collective meaning* is made/refined [20], formalized [6], mediated [17], and applied [4] to form a distributed governance system that closely aligns with Ostrom-ian principles [19].

2.2 Aligning AI Models to Collective Meaning

AI models are increasingly used to address problems of scale in Wikipedia [18]. For example, ORES, an AI model hosting system, is widely used on Wikipedia for a variety of tasks, including identifying damaging edits in articles, assessing article quality, and routing newly created articles to reviewers based on their topics [9].

These AI models are developed to *enact* [12] the artifacts they are supposed to reflect or express, such as the guidelines, policies, and collective meaning of *article quality* on Wikipedia. Recent efforts in AI alignment aim to develop models that ensure an AI's behavior aligns with these artifacts within social and community contexts [7, 21].

In this paper, we argue that—in contrast to the standard problem formulation adopted in AI alignment research—AI models act as *mediators* of collective meaning in a similar way to Wikipedia's policies and guidelines. Like other mediators, AI models "transform, modify, and distort" [15] collective meaning during the translation process. That is to say that "all models are wrong," and achieving perfect, unidirectional AI alignment with collective meanings is impossible [23]. Instead, we will argue that the translation between collective meanings and the application of AI models is naturally *bidirectional* like other mediating genres. In this work, we focus our exploration on this bidirectional relationship between AI models and collective meaning via collective auditing practices. We are not the first to identify the power of participatory AI to encourage

[3]c.f. The "Spirit" of the law: https://enwp.org/Letter_and_spirit_of_the_law

reflection (e.g. [26]), but we are the first to connect this reflective, *collective meaning making* processes and formalization within a genre ecology. We are also the first to observe the structure of this *reversal of the flow of meaning* in situ.

3 Study Method

We adopted a participatory action research approach [5, 13] by working closely with Wikipedia community stakeholders to co-construct research plans and interventions. We initiated the project together with the Dutch Wikipedia community to tackle a challenge they faced. We engaged community stakeholders as co-inquirers and adhered to the community's best practices, for example, by recording our activities with detailed ledgers using wiki pages for documentation and "talk pages" for discussion. We (the developers) drove model design, a periodic formal auditing process, and the development of basic tools for accessing the model predictions in-context. Our community partners drove ongoing informal audits, the design of principles the model was intended to align with, and socialization of the model (understanding, use, and feedback) across the Dutch Wikipedia community. Throughout the project, we co-reflected the research process with community participants [11]. To be clear, this did not start out as a research project. It was merely our attempt at effective-community co-development of a useful technology. However, through this process, we recognized that aspects of our work together provided a new perspective on the discourse around AI alignment, emphasizing the importance of a bidirectional process between AI models and the community's collective meaning. In collaboration with a community partner (the "tool coach", as described later) who served as a co-author, we wrote this paper to share our approach and insights in building AI models, policies, and collective meaning alongside communities.

4 Study Context

In May of 2019, we attended the Wikimedia Hackathon, a yearly in-person event organized by the Wikimedia Foundation that "brings together developers from all around the world to improve the technological infrastructure of Wikipedia and other Wikimedia projects." As part of our activities at that event, we met technically inclined Wikipedians from Dutch Wikipedia who had heard about how article quality models were used in English Wikipedia [8] and were interested in what it might take to set up such a model for Dutch Wikipedia.

We worked together to file a request to build the models in the relevant task tracking system[4] and populated the request with basic questions that are useful for understanding how a community like Dutch Wikipedia already thinks about article quality: e.g., *"How do Dutch Wikipedians label articles by their quality level?"* and *"What levels are there and what processes do they follow when labeling articles for quality?"* We developed these questions through past experiences building similar models for other Wikipedia communities. See Halfaker et al.[8] for a technical discussion of how we developed these models and a key example of their application – tracking the growth of and quality of content at scale in Wikipedia.

The answers to these questions were surprisingly complicated. Wikipedians from the Dutch language Wikipedia reported that they did not have a completed scale. Instead, they had some processes for tagging the lowest quality articles ("Beginnetje") and highest quality articles ("Etalage"), but everything in between had no definition, despite ongoing discussions since 2004[5]. This contrasts to English Wikipedia with levels from Stub, Start, C, B, GA, and FA in ascending order with strict definitions and labeling practices [25]. Participants in the discussion expressed their reluctance to simply adopt the scale from English Wikipedia[6].

At this point, it was clear that setting up an article quality model for Dutch Wikipedia would also require the complicated work of defining a set of guidelines. We therefore followed the mechanisms that Wikipedians use to build consensus and shared understanding about their work. Our Dutch Wikipedian collaborator in May 2020 posted to De Kroeg[7] ("The cafe"), a central discussion space, about the potential of bringing article quality models to the local wiki and included information about how they had been used in other wikis. The proposal was met with light resistance[8] but an agreement was reached that it was acceptable to start experimenting and allow people to use the predictions on an opt-in basis.

Over the next 1.5 years, we engaged in an iterative sensemaking and engineering process using Wikipedians' processes for performing articulation work [24] (or "meaning making" [20]) and their online spaces to co-develop a model and guidelines for assessing article quality in Dutch Wikipedia. Beyond the discussion in De Kroeg, we created an on-wiki project page for the effort[9] where we described the AI model, hosted technical descriptions of the quality scale (see Table 1), posted prediction sets for auditing, and discussed the ongoing work with whoever was interested. Our Dutch Wikipedian co-author gathered a small community of local Wikipedian collaborators around these documents and discussions in order to iterate with us. In the next section we describe aspects of this collaboration that make salient the co-development of collective meaning and AI models.

5 The Case: Dutch Wikipedia Article Quality

5.1 The developer-driven development process

When we first set out to model article quality for Dutch Wikipedians, we wanted to use as much past work as we could before trying to define any new aspects of quality. Dutch Wikipedians had already developed formal processes and definitions for the top and bottom quality classes (beginnetje and etalage respectively). Through discussion with our Wikipedian collaborators, we settled on a rough scale that added three quality levels between these two extremes:

- B-class: Former etalage class articles, and a community compiled list with so-called "rough-diamonds" were assumed to be high quality but not high enough quality.

[4] https://phabricator.wikimedia.org/T223782

[5] https://w.wiki/ANXp

[6] Factors discussed include cultural differences between the English and Dutch language communities and differences article writing guidelines

[7] https://w.wiki/655w

[8] Issues raised include concerns about whether an AI could detect *article quality*, concerns about low quality bots that caused issues recently, fear of losing control over the projects' content to AI, and distrust and push-back on new software developments "imposed" by the WMF in general (e.g., "The MV, German Wikipedia, and Superprotect conflict" - https://w.wiki/A3P7)

[9] https://w.wiki/66qM

- D-class: Articles that were tagged as beginnetje, but this tag was removed later on. We assumed these articles to be slightly higher quality than beginnetje.
- C-class: Articles that were between B- and D-class. We ultimately decided to set a formal length criteria for these articles (between 3000 and 5000 bytes of text).

It was apparent to all involved that this scale was overly simplistic but we suspected that, through exploring the limitations, we might elicit the latent shared understanding [24] of the quality of articles from Dutch Wikipedians. Based on past work in aligning model behavior with communities of Wikipedians [3, 9], we planned to seek feedback and prompt iteration on the quality scale through the auditing process.

5.1.1 The initial audit. The first step in our auditing process involved generating article quality predictions for all articles in Dutch Wikipedia and randomly sampling 20 articles from each predicted class for review (5 classes × 20 predictions = 100 articles in the assessment set). We used article text from the June 2021 database dump of Dutch Wikipedia[10] to generate predictions. Since the quality of Wikipedia articles is highly skewed with the vast majority of articles in the lower quality range, this stratified approach allowed our collaborators to assess the performance across the scale. We posted the list with predictions on a wiki page and invited Dutch Wikipedians to leave open ended comments about each prediction.

Some evaluations implied adjustments to the naive quality scale v1 (Table 1). For example, on an article predicted to be C-class, one Wikipedian commented (translated): "Not a 'good' article, and I would personally rate it as D because of its focus on a summary, and the lack of further sources beyond the one report. But, strictly speaking, does it seem to meet the criteria?" There are many insights in this comment. First, it directly critiques the model's prediction and suggests that the article in question should be rated lower (D-class). It also raises concerns about "focus on a summary" with regards to writing quality, and calls out the lack of sources. It also challenges the naive C-class criteria we started with (3000-5000 characters) as capturing what this editor imagines the C-class should represent.

Many such comments were met with follow-up discussion. For example, another Wikipedian left the following comment on an article predicted to be D-class: *"Only two sources that are both inaccessible, uninformative, clumsily edited. As far as I'm concerned, at the bottom of E."* While a third Wikipedian challenged the downgraded assessment with: *"Please note that E is meant for real beginnetje, unless we find a (measurable) way (and agreement on this) to also include poor quality articles."* In this example, we can see source quality, information quality, and editing quality being raised as important criteria.

Beyond these concerns about the nature of quality and how a scale might be applied to these articles, our collaborators also noted that the model missed critical features of quality such as the presence of Infoboxes[11] and that some *non-articles* were included in the set – such as "list articles", like *Lijst van spelers van Middlesbrough FC* (List of Middlesbrough FC players). We addressed these issues directly through improved feature engineering and sampling methods, which led to AI model v2. Wikipedians produced v2 of the quality scale through this audit-driven meaning-making process.

5.1.2 Labeling and re-auditing. With updated guidelines, we needed new training data that represented the changes. We set out to build a stratified sample of articles to label. Since the consensus on the criteria for the two extreme classes (*beginnetje* and *etalage*) had not changed, we only needed to gather new labels for the middle quality classes (B, C, and D). We applied model v2 to a large set of articles and sampled 25 B-predicted articles, 50 C-predicted articles, and 25 D-predicted articles for labeling. We sampled more C-predicted articles because we expected that the predictions for that quality class were less accurate due to the naive length-based definition and therefore the actual labels for that group would be distributed across B and D classes. In order to ensure consensus on labels, we required three labels per article from different Wikipedians through configuration of the Wiki Labels system [9].

We observed significant disagreement in labeling, with 56 articles showing discrepancies among labelers. Figure 3 shows the first 8 rows of the table we constructed for re-auditing. We went back to the Wikipedians who performed the labeling work and discussed with them about why there might be so much disagreement. Our tool coach started a discussion around the requirement "alles boven E heeft minimaal 1 bron" (everything above E has at least 1 source) Several editors reported that they applied this source criteria very strictly and observed that old styles of sourcing content (e.g. via a comment associated with an edit) could be the reason that some seemingly high quality articles are getting labeled as lower quality. The discussion quickly turns into reflection about what aspects of quality they wish to capture in their scale. For example:

- "Kijkend naar deze uitslagen, denken jullie dat de 'geen bronvoorwaarde' in de C-versie van de kwaliteitsschaal juist is? Of moet deze misschien versoepeld worden?" (Looking at these results, do you think the 'no source condition' [for downgrading from C to E class] in the C version of the quality scale is correct? Or perhaps this should be relaxed?)
- "Van mij mag de grens 'bron/geen bron' wel een niveau hoger" (For me, the threshold 'source/no source' may be set one level higher [i.e. moving it from E to D class])
- "[...] ik denk dat het bron-criterium wel een goede reflectie is van de kwaliteit." (I think that the source criterion is a good reflection of the quality.)
- "Jouw voorstel om de broneis te verplaatsen naar C spreekt me wel aan." (Your proposal to move the citation requirement to C appeals to me.)

Based on this discussion, our Wikipedian collaborators updated the quality scale to reflect the new consensus to move the source requirement to C-class and to soften the language of what can be considered a source ("eventueel als een algemene bron onder een kopje 'literatuur' of 'externe link"' which translates to "possibly as a general source under a heading 'literature' or 'external link"). This resulted in quality scale v3. Through this discussion and re-auditing, we constructed a hybrid dataset with labels aligned with the updated quality scale.

Finally, we used this dataset to train AI model v3. We ended up training the model on 32 examples from each quality class (32 ×

[10] https://dumps.wikimedia.org/
[11] https://enwp.org/H:IB

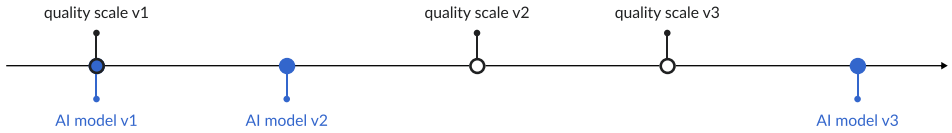

Figure 2: The iterative process of co-developing both the quality scale and AI model over time. Table 1 describes the details of a specific version of the scale or model.

Quality guidelines	AI Model
v1: A and E class are borrowed from pre-defined concepts. B is defined as "not quite A", D is defined as "no longer E" and C is defined by article length.	**v1:** Trained by gathering examples of template introductions/removals and using length constraints. We expected this model to be poorly align but a useful probe for discussion.
v2: B, C, and D classes are more clearly defined. For example, C-class requires the presence of an Infobox and D-class requires that there is at least one source in the article.	**v2:** Trained using the same data as v1 but with minor technical improvements to the way that Infoboxes and references are tracked.
v3: Refined version of v2 based on reflection when applying v2 scale to articles. Source requirement moved C-class and softened.	**v3:** Trained using a mixture of data sourced from template usage (for A- and E-class) as well as the results of labeling activities and on-wiki re-labeling using the v3 quality scale.

Table 1: Alignment between the model versions and quality guidelines

Pagina (versie)	Labels	Def label
Battle Mountain (52460837)	'D', 'D', 'E'	~~E~~ D
Don't Change Your Husband (45963723)	'D', 'E', 'E'	E
Rally van Sardinië 2011 (59732009)	'D', 'E', 'E'	~~E~~ D
Meritsjleri (58471630)	'D', 'E', 'E'	~~E~~ D
Coffee & Co (58779170)	'D', 'E', 'E'	D
Holebikort (59396686)	'C', 'D', 'D'	C
Station Herne (België) (58565985)	'C', 'D', 'D'	D
Ingeborg Sæhlie (42137552)	'D', 'D', 'E'	D

Figure 3: An abridged screenshot of the table we constructed for labeling and re-auditing. The "labels" column represents the original labels that were submitted via the Wiki Labels system. The "def label" column was filled in after a discussion based on consensus.

5 = 160 total articles). Despite this small training set, we achieved 80.8% accuracy across the five quality classes and agreed to deploy the model for testing with our Wikipedian collaborators.

5.2 The Community-Driven Development Process

In parallel with the developer-driven labeling and auditing process, our Dutch Wikipedia collaborators, led by a "tool coach", also drove a complementary effort to support the development of quality scales and AI models.

5.2.1 The tool coach. One co-author of this paper is an administrator and active contributor on Dutch Wikipedia. She offered to work with the community as a "tool coach," a term coined by Sumana Harihareswara to describe someone who fills a bridging role between communities and the technical contributors, helping out in the bits that maintainers are not great at, or don't have time for [10]. When we shared the first version of the article quality model with the Dutch Wikipedia community, she developed an effective strategy for communicating the strange and inconsistent AI behaviors that people would see.

5.2.2 The strange ducks. Our tool coach organized a space for our Dutch Wikipedia collaborators to share any behaviors they thought were wrong, unexpected, or otherwise worthy of discussion. She named this area "vreemde eenden in de bijt" which roughly translates to "strange duck in the pond" – a Dutch euphemism for odd things that don't belong. She then developed a weekly cadence to review the submissions in discussions with Wikipedians on the associated talk page, and brought a summary of those discussions to the development team. As a local community member, she was able to help answer our developers' questions about why a behavior is considered to be strange and what behaviors might be more aligned with Dutch Wikipedians' expectations. She also discussed these issues with submitters in their native language and situated within their shared cultural context.

5.2.3 The gadget in situ. A key to making this community reflection work was getting the model's predictions in front of Wikipedians in the course of their *regular activities* on Wikipedia. To do so, we developed a JavaScript-based gadget that Wikipedians could enable in their Wikipedia account settings. This gadget offers automated article quality predictions while the editor browses and works on Wikipedia, as shown in Figure 4.

Through the gadget in situ and the repository of strange ducks, our tool coach formed a cross-lingual, cross-cultural bridge between the development team and the Wikipedians to support community-driven reflections about the behavior of AI models and the implications of the quality scales. While the developer-driven audits and labeling provided focused opportunities for review and reflection among our Wikipedian collaborators, the community-driven process is more continuous and enables specific concerns to be raised,

Figure 4: (a) Article quality predictions appear on all article pages under the page title. (b) Article quality predictions appear on the revision history page. The colored rectangles to the left of the revision details denotes the prediction. Vandalism that lowers the article's quality is made apparent by a yellow rectangle that shows a temporary drop in predicted quality. (c) Article quality predictions next to article links.

with specific examples, at any point in time. This technology and social process formed a key component of the co-development process – enabling and grounding reflection and renegotiation of the collective meaning.

5.3 Fitting It All Together

Figure 5 shows how the developer- and community-driven processes together enable the co-development of AI models and collective meaning. In particular, labeling disagreements and "strange ducks" are great example of a *high value interaction*. If Wikipedians disagree on a data point or flag it as a strange duck, there are three potential reasons, as shown in Figure 5. Deciding which case the data point fits into is a matter of discussion, but regardless of the results of the discussion, the outcome is valuable to the functioning of the entire system. Either the model needs to change, the guidelines need to change, or the guidelines need to be clarified (or more carefully considered). Each of these represents an opportunity for meaning-making, reflection, and reshaping. Grounding the discussion on the specific examples of "strange ducks" and how they should be labeled seemed to focus the discussion on the *usefulness* of the rules expressed in the guidelines and model behavior.

6 Collective Meaning Cascades but Strange Ducks Swim Upstream

In the case of Dutch Wikipedia's article quality, we can see the behavior of an AI model at the intersection of several different branches of HCI and CSCW scholarship.

The documented norms and practices around article quality assessment translate the collective meaning into a share-able representation for Wikipedians. These documents form a genre ecology that captures the formal and informal concepts used by Wikipedians to articulate (work together) in Wikipedia. From concrete work practices to statements of principle, the entire cascade of meaning is intentionally kept in alignment through Wikipedia's meaning-making processes built on top of peer discussion and documentation practices [20].

The work of developing and refining an AI model extends the rules from the on-wiki text documentation into the behavior of the model itself. In the same way that one might code rules into best practice documentation, one can see rules play out in the AI model's behavior. As Figure 1 suggests, policies represent a way of understanding the collective meaning of Wikipedians, and guidelines represent a way of understanding policies. We assert that AI and machine learning models designed to apply guidelines also represent a way of *understanding* that guideline in a specific setting. As an algorithm, the AI model represents a set of executable rules that can be applied directly, in practice.

All models are wrong [23] is a common aphorism that we find useful when considering the implications of this cascade of meaning. Models, as algorithmic mediators of process, *"enact the objects they are supposed to reflect or express"* but they are inherently imperfect in that they *"transform, translate, distort, and modify the meaning or the elements they are supposed to carry."* [12, 15] Rather than trying to develop a "correct" model, our goal is to design a model that is useful. With an AI model, usefulness is often measured through fitness statistics, but in our case, the collective auditing pattern (e.g. "strange ducks") allowed us to go beyond detecting the error rates and to ask, "How much does this type of error affect the usefulness of the model?" and "Is this an error in the model; is it an opportunity to clarify on the guidelines; or is it an opportunity to re-make collective meaning?" These questions focus issues of alignment on the intended *use* of the model and away from the impossible and less actionable idea of *correctness*.

Further, this is not a special case for AI models. This pattern of error detection, utility assessment, and refinement is consistent across the cascade, from collective meaning to AI model. Just as we can detect modeling bugs through the application of the model and reassessment, we can detect bugs in best practices by exploring whether the model's "bugs" are failures to accurately represent the guideline, whether the guideline fails to accurately represent the policies (the principles behind guidelines), or whether the policy fails to usefully reflect the shared collective meaning of the members

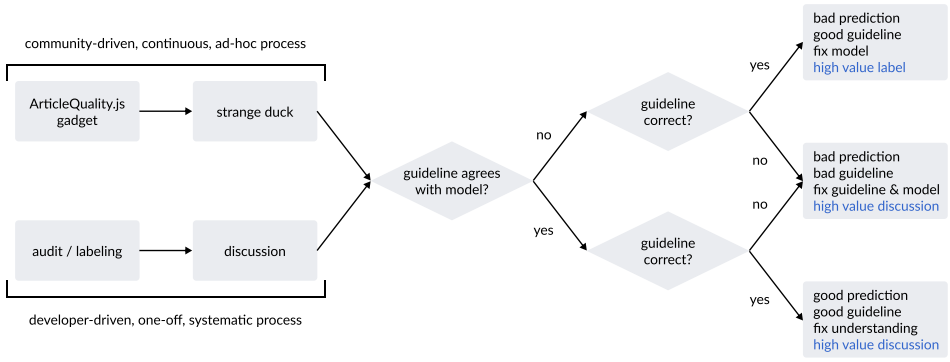

Figure 5: The workflow for developing AI models and collective meanings in our case study. Data points with labeling disagreements or strange model predictions can prompt valuable discussions and lead to iterations of AI models, guidelines, or shared understandings of collective meaning.

of the community. In effect, strange ducks (and other participatory auditing practices) allowed Dutch Wikipedians to *swim upstream* in the cascade and make new collective meaning effectively.

For example, in the case of article citation requirements, the question of whether or not model behavior was a bug or not brought the discussion all of the way up to the level of collective meaning. The original guideline that required at least a single citation for being included in D-class and higher, seemed like a reasonable application of the collective meaning and policy around article quality. But in practice, when reviewing the predictions of a model based on that guideline and labeling new articles, it became clear that this was a misalignment between the collective meaning and practice. Through review and discussion, guidelines were updated to reflect this new shared understanding drawn directly from the context of work. And thus the modeling process was updated and applied in order to ensure that the new model (v3) would reflect this update to the meaning cascade.

7 Discussion & Conclusions

7.1 Models as a Mediator in Participatory Governance

In this work, we observed AI models fill a similar conceptual role to other *mediators* present in Wikipedia's document genre ecology. We find that considering these AI models as a novel genre in the ecology helps us understand the role they play in conveying meaning to practice and opens new doors considering how they fit together with the reflective, norms development and refinement strategies in communities of practice.

One topic that came up early in our work was, *"Why not just adopt English Wikipedia's collective meaning and quality model?"* After all, it was developed for Wikipedia, by Wikipedians. Our Dutch collaborators were very clear that they wished to come to their own definition and have their own model. We see Ostrom's principles playing out in model development the way that they were observed to play out in policy and guidelines development by Forte et al. [6]: that self determination of the community must be recognized and the appropriation and provision of common resources are adapted to local conditions (cf. principle #2 and #7 from [19]). At first, the Dutch Wikipedia community was apprehensive about welcoming AI models into their work. But through our *Ostromian* processes, Dutch Wikipedians were centered in the model development/meaning-making process. And just as Ostrom observed that rules are more likely to be followed by people if they had a hand in writing them, we observe that both models and guidelines are more likely to be appropriated by people if they had a hand in developing them.

7.2 Community Audits as Grounded Reflection on Utility

As we discuss above, Dutch Wikipedians had long struggled with defining "quality." The community had attempted several different initiatives to apply their meaning making process to build shared understanding and identify *collective meaning* around what quality is since 2004 with each effort failing to build agreement. Our community partner (the tool coach) reflects that the iterative process of auditing the AI model and refining the guidelines we describe above kept Dutch Wikipedia participants engaged and more focused on the *utility* of the model/guidelines than on the *correctness* of either. As Sterman observed [23], so do we: *"Because all models are wrong, we reject the notion that models can be validated in the dictionary definition sense of 'establishing truthfulness', instead focusing on creating models that are useful [...] We argue that focusing on the process of modeling [...] speeds learning and leads to better models, better policies, and a greater chance of implementation and system improvement."* In our words, grounding discussion of collective meaning in the *usefulness* of models in practice facilitated the *making* of meaning that was persistently latent despite several attempts to bring about a consensus.

7.3 Generalizability Beyond Wikipedia

As one of the largest, most successful online communities for collective knowledge building, there is much for others to learn from Wikipedia for the development of AI models and collective meaning. Across a wide range of contexts, mediation is a pervasive pattern by which meaning gets applied in practice, whether through AI

models [3, 9], deterministic algorithms [12], or documentary practices [6, 17]. We suggest that efforts in AI alignment [7, 21] should consider the crucial role of mediating artifacts in any context, including AI models as mediators themselves. For example, considering the law as a mediator and the *spirit of the law* as the collective meaning, we encourage developers to promote discussion around how an AI model or algorithm enacting a law also enacts the spirit of that law. Having an AI model applied in practice helps us ground the discussion around utility rather than correctness [23].

Building on this work, future research in AI alignment should develop systems and methods that recognize and support the bidirectional flow of meaning between different layers in the collective meaning cascade framework. This suggests that researchers and practitioners should move away from the notion of aligning to a fixed, pre-existing, mediating documentation. Instead, it may be more productive to embrace a deeply conversational, multi-turn approach to AI alignment, which acknowledges that collective meaning is *actively co-developed* alongside mediating documents and mediating AI models. By grounding discussions of *usefulness*, model co-development processes are an opportunity to make meaning more effectively.

Acknowledgments

The funding for this research was provided by UL Research Institutes through the Center for Advancing Safety of Machine Intelligence, and CMU's Block Center for Technology and Society.

References

[1] B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II 12*. Springer, 277–288.

[2] Sumit Asthana and Aaron Halfaker. 2018. With few eyes, all hoaxes are deep. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.

[3] Sumit Asthana, Sabrina Tobar Thommel, Aaron Lee Halfaker, and Nikola Banovic. 2021. Automatically Labeling Low Quality Content on Wikipedia By Leveraging Patterns in Editing Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 359 (oct 2021), 23 pages. https://doi.org/10.1145/3479503

[4] Ivan Beschastnikh, Travis Kriplean, and David McDonald. 2008. Wikipedian self-governance in action: Motivating the policy lens. In *Proceedings of the International AAAI Conference on Web and Social Media*. 27–35.

[5] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (<conf-loc>, <city>Boston</city>, <state>MA</state>, <country>USA</country>, </conf-loc>) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. https://doi.org/10.1145/3617694.3623261

[6] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems* 26, 1 (2009), 49–72.

[7] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[8] Aaron Halfaker. 2017. Interpolating quality dynamics in Wikipedia and demonstrating the Keilana effect. In *Proceedings of the 13th international symposium on open collaboration*. ACM, 1–9.

[9] Aaron Halfaker and R. Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 148 (oct 2020), 37 pages. https://doi.org/10.1145/3415219

[10] Sumana Harihareswara. 2021. Sidestepping the PR Bottleneck: Four Non-Dev Ways To Support Your Upstreams. https://www.harihareswara.net/posts/2021/sidestepping-the-pr-bottleneck-four-non-dev-ways-to-support-your-upstreams/#coaching-and-cheerleading. Accessed: 2024-05-27.

[11] Dorothy Howard and Lilly Irani. 2019. Ways of Knowing When Research Subjects Care. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300327

[12] Lucas D Introna. 2016. Algorithms, governance, and governmentality: On governing academic writing. *Science, Technology, & Human Values* 41, 1 (2016), 17–49.

[13] Stephen Kemmis, Robin McTaggart, Rhonda Nixon, Stephen Kemmis, Robin McTaggart, and Rhonda Nixon. 2014. Introducing critical participatory action research. *The action research planner: Doing critical participatory action research* (2014), 1–31.

[14] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 193, 24 pages. https://doi.org/10.1145/3613904.3642278

[15] Bruno Latour. 2007. *Reassembling the social: An introduction to actor-network-theory*. Oup Oxford.

[16] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198* (2018).

[17] Jonathan T Morgan and Mark Zachry. 2010. Negotiating with angry mastodons: the wikipedia policy environment as genre ecology. In *Proceedings of the 2010 ACM International Conference on Supporting Group Work*. ACM, 165–168.

[18] Claudia Müller-Birn, Leonhard Dobusch, and James D Herbsleb. 2013. Work-to-rule: the emergence of algorithmic governance in Wikipedia. In *Proceedings of the 6th International Conference on Communities and Technologies*. 80–89.

[19] Elinor Ostrom. 1999. Design principles and threats to sustainable organizations that manage commons. In *Workshop in Political Theory and Policy Analysis, W99-6. Center for the Study of Institutions, Population, and Environmental Change. Indiana University, USA. www. indiana. edu*.

[20] Joseph Michael Reagle. 2010. *Good faith collaboration: The culture of Wikipedia*. MIT press.

[21] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A Roadmap to Pluralistic Alignment. *arXiv preprint arXiv:2402.05070* (2024).

[22] Clay Spinuzzi and Mark Zachry. 2000. Genre ecologies: An open-system approach to understanding and constructing documentation. *ACM Journal of Computer Documentation (JCD)* 24, 3 (2000), 169–181.

[23] John D Sterman. 2002. All models are wrong: reflections on becoming a systems scientist. *System Dynamics Review: The Journal of the System Dynamics Society* 18, 4 (2002), 501–531.

[24] Lucy A. Suchman. 1994. Suporting Articulation Work: Aspects of a Feminist Practice of Technology Production. In *Proceedings of the IFIP TC9/WG9.1 Fifth International Conference on Woman, Work and Computerization: Breaking Old Boundaries - Building New Forms*. Elsevier Science Inc., USA, 7–21.

[25] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell me more: an actionable quality model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*. 1–10.

[26] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.