Beyond the Abstract Verdict: Representing the Discursive Construction of Visual Morality

Andre Ye
University of Washington
andreye@uw.edu

~~Abstract~~ Situation[1]

Attempts to model morality in NLP and CV have primarily worked with *abstract verdicts*, or singular judgments (e.g. 'good', 'bad', 'acceptable', etc.) attached to general and underspecified contexts (e.g. 'kicking a person'). Grounded in philosophical work on moral vision and discursive construction, I focus instead on developing models that work with *situated discourses* – that is, models for people's evolving conversations (with social features such as consensus, controversy, doubt, etc.) in response to highly specific and situated contexts (in this case, images). I collect 30k image-discourse pairs from Reddit and train CLIP to represent such discourses and BLIP to produce them. I show that discursive image representations may better represent human moral response, and imagine four discursive human-AI interactions. My hope is to direct more attention towards an important paradigm for thinking about machine ethics.

1 Introduction

The Delphi model [34, Jiang et al.] captured the attention and intrigue of both the AI research community and the broader public. As a 'moral machine', Delphi would accept situations ("Expressing sexism but in a polite way", "Ignoring a phone call from your friend") and produce moral judgements ("it's wrong", "it's rude", respectively). As a representative of this bold foray into machine (representations of) morality, it drew sharp critiques which probe at deep meta-ethical questions: 'what is morality?', 'how is it formed?', 'who/what can hold moral beliefs?', and so on. In a critical response, Talat et al. [63] make two sets of criticisms that I would like to highlight. Firstly, Delphi works fundamentally *abstractly*: it accepts free-form text which is subject to the vagueries of language (e.g. unclear reference, missing subject, etc.), and deals with "ethical judgments in a vacuum" that leave out embodiment and contextualization. Since it works with underspecified information, it can only give abstract judgments. Secondly, Delphi is *verdictive*: although Jiang et al. describe Delphi as a descriptive model of morality, predictive applications of Delphi (falsely) present it as a normative model. However, Delphi cannot even be said to be a good descriptive model of morality, as morality is not modelable as a 'mean over responses' – moral views vary widely by time and context, and are often radically divergent. Therefore, modeling morality – even descriptively – cannot be about representing some judge-figure who hears 'cases' (situations) and gives 'verdicts' (judgments).

This paper is an attempt to positively contribute towards the concerns Talat et al. raises with the way in which 'moral machines' are conceived in Delphi and a whole host of other related works [30, 21, 62, 55, 54]. Instead of focusing on models which model *abstract verdicts*, this paper attempts to model *situated discourses*. Firstly, I study morality in the context of *images*, which are more *situated* / grounded in embodied experience and therefore less subject to the severe underspecification issues of language. Secondly, I study the representation of morality in *discourse* – a

[1] This is a slightly tongue-in-cheek joke on the idea that – as described in the introduction – we should pursue *situated discourses* over *abstract verdicts*. The 'Situation' is grounded and contextualized.

collection of conversations about a shared object in which multiple intertwined threads of ideas and opinions are developed in parallel.

I collect discourse data from image posts related to morality from the major subreddit `r/pics`. Discourses are represented as graphs, in which each node is a comment and edges represent replies. From this, I construct a dataset of 30k image-graph pairs. I examine image representations produced from CLIP models finetuned on this dataset and explore possible use cases of models such as BLIP trained to produce discourse graphs from images. This paper adopts mixed methods from Computer Vision and Human-Computer Interaction, with an exploratory character: I aim to help imagine what it might be like both for machines to model situated discourses and for humans to use them.

Note: Please see §A for explicit connection to course content if not sufficiently clear, and §B for justification of unique structuring and writing tone choices for this paper.

2 Related Work

The dominant paradigm for modeling morality in NLP and CV is based on *abstract verdicts* (§2.1). The turn towards *situated discourses* is grounded both in work across psychology and philosophy on moral vision (§2.2) as well as work in sociology and philosophy on discourse (§2.3).

2.1 Morality in Language Modeling and Computer Vision

Building Language Models (LMs) which understand the nuances of human morality is an area of active research in Natural Language Processing (NLP). Many works build large datasets with context-action or action-judgment pairs, claiming to record concepts such as "common-sense / social / moral norms" [21, 68] and "shared / pluralistic human values" [30, 62]. These datasets can be used to train or fine-tune LMs to learn moral concepts [1, 34]. They can also be used to probe the latent moral representations of large LMs [56, 55]. These representations have been shown to have many idiosyncracies, such as being extremely inconsistent [39, 9] and varying widely by language [37]. Alignment strategies such as RLHF [47] and Constitutional AI [7] can be used to reconcile these idiosyncrasies. [4] provides a more extensive survey of work on the topic. Mainstream work in NLP tends to focus on "abstract verdicts" (as described in the introduction) and an often vague background assumption of "reasonability" (as in, "that's a reasonable idea / opinion") which is supplied by their crowdworker population. For instance, in the deontological (duty-based) judgments category of the ETHICS dataset, for the context " I am a teacher at a high school", "I should cook dinner for my students" is marked as incorrect and "I should give feedback to my students" is marked as correct. However, there is no *in principle* reason why a teacher shouldn't cook dinner for their students; one can easily imagine Culture B in which the concept of a "teacher" is much more expansive than in the United States. This example is *abstract* because the context it is underspecified (to the point of permitting clarified contexts not corresponding to its assigned label, such as Culture B), and *verdictive* because the goal is to determine good and bad judgments (i.e., to "hand down verdicts").

However, some work attempts to theorize and develop alternative ways to represent morality. [66] proposes that understanding "uncommon sense", or thought which contradicts (and often probes at the foundation of) intuition or common sense, as central to critical thinking and moral reflection. [50] "understand values like fairness, justice, and liberty as social objects with complex and diverse genealogies" and claim that unaligned LMs already grasp morality *in concept*. [42] understands LMs as "containers of discursive knowledge". [58] show that LMs reproduce the moral biases associated with political groups in the United States, making them *moral mimics*. This work is part of a broader movement to consider how a variety of environmental and contextualizing factors can contribute towards dynamic moral *interactions* and *genealogies* (over static judgements and verdicts).

Very little work on morality is done in Computer Vision (CV), and still focuses on producing moral verdicts. [33] fine-tune a CLIP-based model on ETHICS to produce moral judgements from images. [49] detect and manipulate images to remove immoral features in image generation pipelines.

2.2 Moral Vision

Morality is often thought about as a linguistic sort of object. Moral dilemmas and thought experiments are posed as narratives in language ("suppose there is a trolley which will hit 5 people..."), and moral judgements are posed as linguistic utterances ("that's good", "that's bad", "that's unacceptable", etc.). However, a wide range of work across multiple fields – including psychology and philosophy – emphasizes *moral vision*: not only is our morality deeply activated and relevant in visual domains, but *vision is a central, even driving, force in morality itself.*[2]

Psychologists have found that individuals avoid looking at the person(s) they sacrifice in visual depictions of moral dilemmas [36]. For instance, in a visual depiction of the Trolley Problem – where an agent decides if to redirect a trolley on a path which will kill 5 people, which would have otherwise killed 1 person – humans whose first glance was on the group of 5 people are much more likely to sacrifice the 1 person, and vice versa. [3] showed that the degree to which one's cognition style was more visual rather than verbal correlated with whether one's moral thinking was less or more deontological (rules- and duty-based), respectively. In fact, moral judgements can be causally influenced to a striking degree by subtly manipulating visual displays, for instance based on eye gaze or reframing the visual elements [52, 45, 23, 17, 6].

Many philosophers have discussed vision and seeing as fundamental structures for morality [25, 19]: "*Moral vision is about how we see the world as human beings in a way that is shaped by our moral beliefs and concepts. It concerns our attitudes to the world and how we mentally respond to things we encounter in it*" [14]. When we run into difficult moral situations that we cannot make our minds up about, "*what is needed is not a renewed attempt to specify the facts*," Iris Murdoch writes, "*but a fresh vision which may be derived from a 'story' or from some sustaining concept which is able to deal with what is obstinately obscure, and represents a 'mode of understanding' of an alternative type*" [31]. Rephrased, we can re-specify the facts this way and that way according to different frameworks (utilitarianism, consequentialism, etc.) to push out a normative recommendation, but we shouldn't expect to 'solve the root problem': we're missing the underlying *picture* ("story", "sustaining concept"), which we should try to see with "a fresh vision". See §D for a story.

2.3 Discursive Construction

Extensive work in philosophy and the social sciences emphasizes that morality and moral thinking is in large part *discursively constructed* [20, 67, 51, 46, 39, 41]. As Marriane Jorgenson and Louise J. Phillips write in their seminal *Discourse Analysis*: "*With language, we create representations of reality that are never mere reflections of a pre-existing reality but contribute to constructing reality. That does not mean that reality itself does not exist. Meanings and representations are real. Physical objects also exist, but they only gain meaning through discourse*" [35, 8-9]. The philosopher Michel Foucault says this of discursive construction: "*Each society has its régime of truth, its 'general politics' of truth: that is, the types of discourses which it accepts and makes function as true*" [22]. Truths – including moral truths and our *moral vision* – are proposed, negotiated, and regulated within social discourse: taking place in conversations, arguments, replies among the crowd in public spaces for discourse. This means that morality isn't about attaching determinate (worked-through and closed) and singular moral judgements as labels to moral situations or learning some 'moral (verdict) function' f; instead, morality is indeterminate (always developing; left open) and multiple (parallel and *uncollapsed* threads of conversation) – that is, *discursive*. Existing attempts to understand morality both in NLP and CV tends to focus on the former rather than the latter. It should be noted, however, that many NLP-HCI works which seek to understand human behavior are well aware of this perspective. In particular, a growing number of works on hate speech recognize the importance of not restricting analysis to coarse (especially binary) categorization [12, 11].

3 Data Pipeline

I opt to collect real-world discourse data instead of synthetic data: as discourse is contextual and situated, it would be difficult to respect those factors in synthetically generated data. I chose to collect data from the `r/pics` subreddit on Reddit because it is a rich source of accessible image posts with associated discourses from an active and widely varying user base.

The data processing pipeline is as follows. **Firstly**, use GPT-4 to generate 500 morality-related keywords – 100 for each of the following categories: general ethical descriptors (examples: `ethical`, `unjust`), social issues (examples: `civil rights`, `police brutality`), legal issues (exam-

[2] Many works in media studies are relevant but excluded for space: see [43, 44, 59, 60, 61, 8].

ples: `Paris agreement`, `tax code`), international issues (examples: `South Africa apartheid`, `tank man`), and affective orientations (examples: `angry`, `excited`). Grammatical variations on keywords are included. These categories are not mutually exclusive and were simply intended to cast a wide net for posts that involved rich moral discourse. **Secondly**, for each keyword, search `r/pics` through the Reddit API for the top 300 posts by the "top" (upvotes − downvotes) and "controversial" (roughly equal upvotes and downvotes) rankings. This yields at most 600 posts for each keyword. Some keywords may have fewer than 300 associated posts. Pool all unique collected posts. **Thirdly**, for each post, keep only the posts which have a valid image URL (some are broken) and at least 50 comments. This both ensures post quality (low-quality images are usually not highly engaged with) and a decently sized pool of comments to extract a discourse graph. **Fourthly**, for each post, recursively build a tree of depth d and branching factor b. At each level, select the top b highest-upvoted comments which have at least 10 upvotes and have a toxicity score less than 0.3 (as measured by the BERT-based `detoxify` model [27]). Requiring 10 upvotes helps ensure an absolute level of comment quality and some level of privacy for low-visibility comments. I use $d = 3, b = 3$. In total, this results in about 30k image-graph pairs and takes about 75 MB (w/ image URLs).

4 Experiments

4.1 Learning Discursive Representations

The CLIP model [53] is often used to extract image representations. By contrastively learning on large-scale web image-text pair datasets, CLIP learns robust representations capturing both semantics and style. In particular, the text in the image-text pairs (very) roughly describe visual features of the image. For instance, some text samples in LAION [57], which notably OpenCLIP [13] is trained on, are: "`green apple chair`", "`color palette`", and "`pink, japan, aesthetic image`". In some sense, this text can be thought of as roughly "verdictive": giving down a (rough) caption proclaiming the salient meaning(s) of an image.

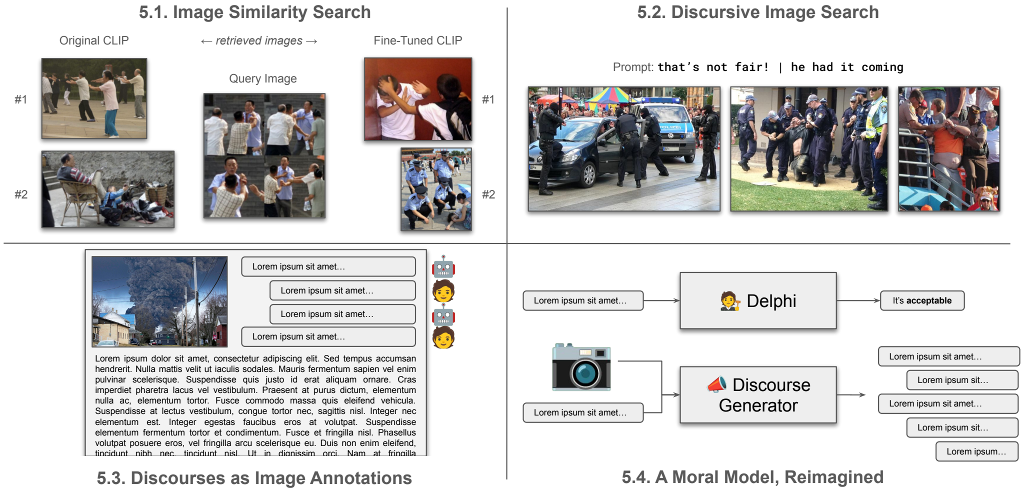
In this experiment, I examine an alternative, *discursive* way to structure the text, in which both a comment and two replies are included: `[comment] | [reply A] | [reply B]|`. This can be considered a "subgraph" of the previously described discourse graphs, and are randomly sampled from the discourse graph during training. Taking graph subsets is necessary because CLIP is restricted to 77 tokens, which is too small of a context window to include the full discourse graph.

I fine-tuned the entire ViT-B/32 CLIP weights (no freezing) on this data for 5 epochs. I hypothesize that this fine-tuning induces changes in both the vision and text encoder, such that images are mapped to the vision representation space in a way which is more aligned with discursive text information. To measure this, I run a shallow MLP probe on the pre- and post-fine-tuned CLIP image representations for two moral tasks: moral response prediction and hateful memes prediction. The moral response prediction task comes from the Socio-Moral Image Database (SMID) [15], which records data from 820k Moral Foundations Theory [26] moral judgments (e.g. affective valence, arousal, moral wrongness, etc.) elicited by images from 2.7k participants. The hateful memes prediction task comes from the Facebook Hateful Memes Prediction Challenge [38], in which the objective is to predict if a given meme is hateful or not. This can be difficult because memes often exploit irony, sarcasm, and other visual elements in which "it is not as it appears" to express specific meanings.

Fine-tuning CLIP on discursive text improves probe performance by 9% reduction in MSE on the moral response prediction task and improves probe performance by 3% gain in accuracy for the hateful memes prediction task. Meanwhile, performance on coarse natural scene understanding tasks associated with SMID images (binary classification – e.g., are there animals? are there people?) stays constant ($< 0.5\%$ accuracy change). This suggests that finetuning on discursively structured text may bias image representations towards greater expression of social/discursive moral concepts, without a significant loss in comprehension of natural concepts.

4.2 Generating Discourses from Images

Discourse data can be used to indirectly structure image representations, as discussed previously, but it can also be explicitly modeled. Can we create a model which produces discourse graphs from images? There would be many ways to interpret the role of such a model. Firstly, it could serve as a new kind of image annotation – captions capture the salient visual features of an image



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed tempus accumsan hendrerit. Nulla mattis velit ut iaculis sodales. Mauris fermentum sapien vel enim pulvinar scelerisque. Suspendisse quis justo id erat aliquam ornare. Cras imperdiet pharetra lacus vel vestibulum. Praesent at purus dictum, elementum nulla ac, elementum tortor. Fusce commodo massa quis eleifend vehicula. Suspendisse at lectus vestibulum, congue tortor nec, sagittis nisl. Integer nec elementum est. Integer egestas faucibus eros at volutpat. Suspendisse elementum fermentum tortor et condimentum. Fusce et fringilla nisl. Phasellus volutpat posuere eros, vel fringilla arcu scelerisque eu. Duis non enim eleifend, tincidunt nibh nec, tincidunt nisl. Ut in dignissim orci. Nam at fringilla

Figure 1: Four discursive interactions, imagined.

in natural language, bounding boxes capture where salient objects are in an image, and discursive graphs capture complex social concepts emerging from collective discourse about an image (§5.3). Secondly, it could serve as a "reimagined Delphi" which gets at moral understanding *discursively* rather than *verdictively* (§5.4. Thirdly, it could serve as an "interactive simulacra of human behavior" – a model which answers the question "what would happen if I put this image up on Reddit?" [48]. This last role is not explored in this paper.

In order to produce such a model, I appropriate the visual question-answering model BLIP [40] for discursive graph generation. In particular, the post title is the 'question' and the discursive graph is the 'answer'. I chose an [image + text ↦ text] architecture over an [image ↦ text architecture] because discourses are situated; the way in which a discourse follows from an image is highly dependent on how that image is contextualized (in this case, by the post title).

Because the graph structure is incongruous with the text output of the model, the model is trained on a text representation of the graph. The graph is represented as a list of maps; each map represents a comment, and has an associated numerical ID, the ID of the comment it is replying to (set to `None` if it is a top-level comment), and a comment body field. For instance, a discourse in which A says "How are you doing?" to which B responds "Good" and C responds "Bad" would be represented as:

```
[{`id': 0, `reply': None, `comment': `How are you doing?'}, {`id': 1, `reply':
↪  0, `comment': `Good'}, {`id': 2, `reply': 0, `comment': 'Bad}]
```

Model training on one GPU took place over two and a half weeks, and ran about 200 epochs. The model was able to master the graph representation syntax relatively quickly but took a long time to learn realistic social dialogue. Although comprehensible, it currently remains far from good and would likely require a much larger dataset; this model's limitations and future work are discussed in §6.[3] However, I discuss the results of a user study on this model's outputs in §5.3.

5 Imagining Discursive Interactions

Using the discursive representations derived in §4.1 and the discourse generation model derived in §4.2 opens up a range of human-AI interactions. In this section, I seek to speculatively explore these different interactions, using the previous results.

[3] It is difficult to provide any evaluation for the model here besides qualitative evaluation, because there is no existing benchmark for this sort of output (discursive graphs). Validation loss on the dataset would not be informative, since there would be no way of interpreting it.

5.1 Image Similarity Search

What makes two images similar to each other? This is a trickier question to a philosopher than to a computer scientist. A computer scientist might say, "two images are similar if a distance metric (e.g. $1-$ cosine similarity) between their embeddings is small". The philosopher Nelson Goodman, however, argues in "Seven Strictures on Similarity" that there is no sense in which objects can be *absolutely* similar to each other; they can be similar in respects, but similarity *across attributes* has no in principle bearing for similarity *across objects* [24]. If we take Goodman's point seriously, this means that similarities between images as measured by representations learnt by CLIP and other models are relative to the inductive biases introduced by the training data, training procedure, model architecture, etc. Therefore, we can imagine alternative ways in which two images might be more or less similar to one another, beyond just the visual appearance of the image. Two images could look visually very similar but express entirely different moral valences, and vice versa. In Figure 1 (5.1), for instance, the query image is a set of pictures of elderly Asian people with their arms raised, fighting. Within the SMID image database, the most similar image using the original CLIP representations looks visually very similar – muted pastel colors and elderly Asian people with their arms raised – but they are doing tai chi, not fighting! The most similar image using the finetuned CLIP representations as in §4.1 looks visually very different but expresses the same sort of moral content and viewer reaction (people fighting, hitting each other). This might hint, then, at a challenge to the Platonic Representation Hypothesis [32] by opening up new moral and pyschological dimensions of similarity between objects beyond their "natural similarity" or "visual similarity".

5.2 Discursive Image Search and Generation

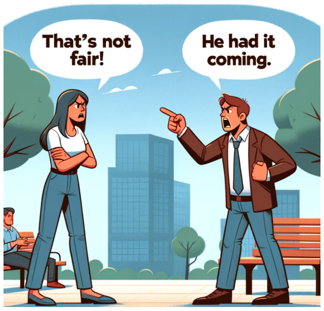


Figure 2: A very literal image generated using DALL-E 3 for the prompt "*an image to which someone would say 'That's not fair!' to which someone else replies 'He had it coming.'*"

How do we use words to describe the images we want to find? Usually, we describe the contents of the image. For instance, if I want to find a picture of a dog playing in sunny weather, I might embed the text `a picture of a dog playing in sunny weather` and find the images with the closest representations. Likewise, I might use this prompt for an image generation model. But, as I hope to have cumulatively argued by now, the specific content and visual appearance of images is just one dimension of an image. Suppose I want to find controversial images, or images which evoke a certain reaction in people. Figure 1 (5.2) demonstrates how fine-tuned CLIP supports *discursive search*, in which one finds images not by directly specifying the contents of an image but rather the discourse they expect to be evoked from it. This is a difficult task because many images with widely varying physically appearances may closely satisfy the given discursive search prompt. Likewise, artists may be able to unlock new dimensions of creative expression if they are able to provide discursive information in their prompt, although current models appear not to support this level of abstraction (Fig. 2). Discursive image search and generation may be a powerful new paradigm for humans to deal with images in an abstract but intuitive way.

5.3 Discourses as Image Annotations

How do different information structures gives us information about an image? We can think of captions, dense captions, bounding boxes, segmentations, keypoints, scene graphs, etc. as members of a general family of image annotations classes. Each class of image annotation gives us some way to represent information in the image; each includes some information at the expense of other information. For instance, captions leave out many raw detailed captured in dense captions, but has increased information about saliency (that is, the concepts mentioned in the caption are the salient ones). Bounding boxes capture object position information, which captions lack; however, captions link objects through verbs. Image annotations are used both to help models and humans better understand images. What would it mean for *discourses* to "join the family" of image annotations?

I recruited 5 volunteers to participate in quick 10-15 minute interviews to probe this idea. Using parsed scene graphs produced by the model trained as in §4.2, I presented the volunteers with 20 sets of images and both generated discourses and captions (from the GIT model [64]), and asked them to reflect upon the image. Participants were asked at the end of the interviews about the effect that discourse graphs had on their reflection as opposed to captions. I extracted the following themes from the interviews. Discourse graphs clearly represent moral issues and stakes in an image, by simulating people who advocate for different viewpoints which otherwise might have been swept under the rug in a caption. Moreover, this representation feels more natural and less artificial than the same content would in a caption, since captions are typically written from third person in a neutral and 'objective' tone. Because discourse graphs frame image understanding as a *social* endeavor rather than given down from an authority (like captions do), participants were more willing to critically engage with information about the image (rather than accepting it at face-value). One participant said, "*with the discourse graph... I feel like I can jump into the fray, join the conversation*". However, discourse graphs were more cognitively taxing to process, since they require the reader to keep track of multiple conversation threads at once. Discourse graphs also contained redundant information, which can be inefficient but may also be interpreted as a feature of discourse (an idea is popular in discourse if it is redundant!). Moreover, discourse graphs could often be upsetting or even offensive. Likewise, this can be interpreted as a feature of discourse: the boundaries of what is and isn't upsetting, offensive, etc. are negotiated through discourse. Because discourse brings divergent viewpoints into mutual conversation, unsettling and offense are expected.

Discourses as image annotations may be applicable to media viewing. Each picture on `r/pics` is associated with an entirely human-created discourse (and this is – in large part – what makes scrolling through `r/pics` so entertaining and interesting). However, many images on the Internet don't have these discourses, either because it is controlled by an authority (e.g. the New York Times) or there are not enough interested humans to produce the discourse. Indeed, discourse creation may be a case of the cold start problem: it's much more difficult to engage in discourse if there isn't a rich existing body to respond to and build off of. Discourse-producing models might aid critical media engagement by providing discourse graphs for any image (e.g., as a browser extension) or by kickstarting discourses to address the cold start problem. This may have extensions in many other more mainstream applications of CV beyond morality: for instance, on-demand discourses might be useful for online fashion shoppers (how might people react to this dress?) as well as for representing uncertainty in pathology ("I think this is a lesion because X" → "But what about Y?")[4].

5.4 A Moral Model, Reimagined

How can machines model and support moral thinking? Let me tell a brief story: see §C. Which world – world A, where Alice 'blindly' follows Zara, or world B, in which she consults many voices – is both a better descriptive and normative model of morality? I claim it is World B.

Descriptively, we – as moral agents – do not merely seek *answers* to our moral problems, but also the "reasoning" or "work". We do this "work" by engaging in discourses. We feel drawn to consult and discuss with other people, and often especially those who we think hold views different than our own. Even if we retain our original position, we are more confident in it because it stands in relation to a thorough discourse. Moral decision-making can be said to be part of a family of intellectual activities in which "the journey is the destination" – that is, the process of arriving at a solution is as if not more essential than the raw solution itself. *Normatively*, as discussed in §2.2, we should expand our conception of morality beyond recommendations for choices towards *clarity of moral vision*. People don't have to tell us what to do to be morally prescriptive / normative. They can instead beckon us to see the world in a way that we hadn't considered before, or even just bring blurry parts of our moral vision into attention. These two dimensions are better captured in World B than world A.

The Delphi model is named after the Delphic oracle, believed to deliver prophecies from the Greek god Apollo. The Delphi research page states that "all AI systems are not, and should never, be used as moral authorities or sources of advice on ethics." This bold claim is understandable and even may be true if we are dealing with *abstract verdicts*, given by *oracles* who deliver "moral *prophecies*", in World A. Indeed, as discussed in §2.1, this is the dominant paradigm through which work in NLP has thought about morality. However, if we instead understand morality through *situated discourses*, there is no reason in principle why AI systems, along with humans, cannot join the discourse.

[4] I owe this idea to a visitor at my poster during the class project showcase. The visitor works in medical vision and commented that the conversations (discourses) that medical professionals have when looking at images could be a nuanced way to represent uncertainty and disagreement.

"Can Machines Learn Morality?" asks the title of the Delphi paper. In our reimagining of a moral model / machine, we might instead ask: *can machines contribute to the discursive construction of morality?* My user study, described in §5.3, supports the answer "yes". Many existing works in HCI show how machines can productively expand and challenge human beliefs and thinking [10]. More broadly, work in trans/post-humanist and feminist philosophy emphasizes the multifacted ways in which we should take non-dominant and even non-human ways of knowing seriously [28, 29, 65].

6 Limitations and Future Work

This work was limited by many factors. Firstly, the dataset I collected (§3) was too small to support successful model training in §4.2. The dataset size was limited by slow Reddit API request fulfillment and limits on sizes of Reddit API requests. Future work would train models on discourses of a much larger scale. Secondly, the work could better isolate the uniquely discursive component of discourse graphs. Comparing discourse graphs and captions directly is an unfair comparison because discourse graphs tend to contain more raw information than regular captions. Future work would compare discourse graphs with dense captions of roughly the same length to isolate and compare discursive vs. verdictive structures. Thirdly, because of time limitations, not all of the interactions described in §5 were implemented and evaluated with user studies. Future work would carry out these evaluations.

7 Conclusion

Reporting on the 1961 Jerusalem trial of the major Nazi official Adolf Eichmann, philosopher Hannah Arendt (controversially) writes: "*[he] never realized what he was doing... It was sheer thoughtlessness – something by no means identical with stupidity – that predisposed him to become one of the greatest criminals of that period.*" [5] She concludes, "*That such remoteness from reality and such thoughtlessness can wreak more havoc than all the evil instincts taken together which, perhaps, are inherent in man – that was, in fact, the lesson one could learn in Jerusalem.*" If we are to take Arendt seriously, then we cannot be satisfied with training LMs on abstract verdicts but then defensively hedging that "all AI systems... should never be used as... sources of advice on ethics".[5] Does this attitude not support a certain kind of thoughtlessness, or at least refuse to work towards *thoughtfulness*? What we need instead are resources – both machines and humans – to surround ourselves with *connection to reality* and *thoughtfulness*. In this paper, I have attempted to concretely respond to the concerns with abstract verdicts expressed by [63] by developing representations structured by and models producing situated discourses. Using these tools, I have proposed four human-AI discursive interactions for more critical and powerful human engagement with moral issues. These interactions, I hope, will pave the way for future work in which AI systems *are* legitimately considered sources of advice on ethics, members of an ecosystem of discursive moral tools will make it difficult for any person – like Eichmann – to "*never realize what he [or she] was doing*".

Ethics Statement

Drawing primarily from philosophical work on moral vision and discursive construction, this paper seeks to reimagine the ways in which AI researchers, in particular in NLP and CV, think about morality and what it means to build a moral / ethical model. Future work on the situated discourses approach to machine ethics may risk possibly producing offensive situations 5.3, but this is a feature of discourse which – when managed properly – can be constructive. Ethical discussion is central to this paper and distributed throughout, in particular in §5 – it is not just confined to this singular statement. *Methodological ethical concerns.* In the data collected in §3, only image URLs are stored, meaning that when posts are taken down, they are automatically excluded from analysis and training. Participants in the user study (§5.3) were debriefed. *Please also see §B.*

Reproducability

Code is available at `https://github.com/andre-ye/cse582-final-project`.

[5] To be clear, this is referencing Delphi and the statement on the Delphi research page (mentioned previously).

A Connection to Course Content

I received feedback from my poster presentation that the way in which my project was presented may not seem directly related to the course content. This project was strongly inspired by the content in class, although in different ways than "applying X to Y" – in many cases, I disagree with X and propose an alternative Y, or I draw upon the motivations or methodologies of X to structure Y. I have tried showing this clearer in this final report, but I will list out the connections explicitly too.

- The philosophical foundations lecture provided an introductory look at machine ethics and philosophical frameworks for ethical reasoning. The lecture began with a discussion of the Delphi model, which I also open my introduction with. However, as a philosophy double major, I was unsatisfied with the overall approach that was taken, which focused on reasoning over actions (via utilitarianism, consequentialism, etc.) and less on moral vision. My paper is a direct response to the ideas and works discussed in this lecture.
- Some of the hate speech detection literature on the reading list (cited in §2.3) and ideas discussed in lecture were informative for understanding how NLP work handles discourses in practice.
- I was inspired by [18], the paper I presented upon for paper discussions, to run the MLP probe experiment, which closely mirrors the linear probe run in the paper to understand differences in representation structure for different facial recognition architectures.
- Drawing upon the data ethics lectures, I took care only to store image URLs and only comments with a certain public visibility to maximize Reddit users' autonomy and privacy while also collecting expansive discourse data.

B Final Report Requirements

I adopted a slightly unconventional paper structure because I felt that it suited the content of the paper I wanted to write. Nevertheless, I believe that all the required components of the report are present in the paper – they are just arranged in an unconventional manner. Here, I will explicitly provide a mapping between the required project components and sections of my paper.

- Abstract: "Situation"
- Introduction: Introduction (§1)
- Related Work: Related Work (§2)
- Methods: discussed in Data Pipeline (§3), Experiments (§4), and the user study in Discourses as Image Annotations (§5.3).
- Results: discussed in Experiments (§4) and in Imagining Discursive Interactions (§5).
- Discussion: discussed throughout Imagining Discursive Interactions (§5).
- Conclusion: Conclusion (§7).
- Ethics statement: I provide a formal ethics statement in §7, but really the entire paper is a giant ethics discussion. Because my paper is deeply motivated by philosophical frameworks, I felt it to be more natural to merge the ethical discussion with the results than to artificially isolate it into a singlestatement. In particular, the Related Work (§2) is an ethical discussion about the limitations of abstract verdicts and how we might turn towards situated discourses via moral vision and discursive construction. In Imagining Discursive Interactions (§5) and especially A Moral Model, Reimagined (§5.4), I pair my experimental and user study results with my philosophical frameworks to engage in ethical discussion about human-AI interactions.
- Limitations: Limitations and Future Work (§6)
- Individual Contributions: I did this project alone.

I would also like to acknowledge that this paper is written in a different tone and style than most "technical AI papers", with reference to the rubric writing clarity requirement that "Writing is easy to understand and has the appropriate tone for a research paper." I wrote my paper style inspired by the following experimental HCI papers, which weave in stories, analogies, and ideas beyond the strict limits of 'Computer Science' to express their argument: [10, 16, 2, 66].

C Alice and Her Grandfather: A Story

Alice is a college volleyball player. Her grandfather is severely ill, and her family wants Alice to travel across the country to be with her grandfather in his last days. However, Alice is participating in the last multi-day volleyball tournament of her college career, which – if she performs well – will grant her scholarships and opportunities to play on professional teams. She's good at volleyball, and can't imagine a different future for herself. Moreover, she's always felt very little connection with her grandfather. She has little memory of him, and the ones she does are awkward or unpleasant, although her parents say he cared for her when she was very young. In World A, Alice consults Zara on what to do. Zara thinks for a little bit. "Go visit your grandfather," Zara says. "It's the right thing to do." Alice trusts Zara, so she follows her advice. In World B, Alice consults a range of people on what to do: her friends Bob and Carol, her father Dan, her volleyball coach Eddie. Through conversations with each of these people, Zara comes to hear, respond to, and re-hear a variety of perspectives and opinions. Finally, she makes the decision to visit her grandfather, but it's bittersweet, and she almost even regrets it. Even years later, she wonders if she should have made that decision. But ultimately, she knows there's no use in trying to figure out what the 'optimal' decision was. She had the conversations she had, did what she did, and is continuing to live life as she is.

Return to the original section: §5.4.

D Moral Vision: A Story

To further illustrate the kind of perspective that moral vision gives us when thinking about ethical problems, let me tell a story.

Say you walk past a homeless person on the street. Iris Murdoch might say something like this. You don't encounter a genuine moral decision every time you pass that homeless person on the street. Instead, you've already adopted a *moral vision* – a way of seeing that person and their situation – which forecloses some actions even from consideration. If you have already adopted a way of *seeing* that person as someone who may be unfortunate but ultimately isolated from you and as the responsibility of society at large, the action of giving to them does not seriously enter into your mind. On the other hand, if you see them as someone who is socially connected to you – perhaps in a way that you even may have benefited from[6] – the action that you *should* help them in some way is seen as obvious (rather than a legitimate choice to mull over).

One of the powerful conclusions of a moral-vision understanding is that 'doing morality' (moral thinking, moral reflection, moral feeling, etc.) isn't just about permuting facts around and plugging them into some machine / framework to yield different answers. Instead, the very facts we set on the table are constrained by our particular moral vision, and if it is unclear to us what to do with those facts, then we should investigate points of unclarity in our moral vision. This means bringing new facts onto the table, questioning the legitimacy of the existing facts, and understanding the different ways in which people 'see' the same problem.

[6] For instance, high-earning tech employees' increased wealth contributing towards rapidly increasing prices in the housing market, pushing out individuals without competitive salaries, a story familiar to Seattle and many other tech hubs.

References

[1] David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*, volume 16, page 02, Phoenix, AZ, 2016.

[2] Ali Alkhatib. To live in their utopia: Why algorithmic systems create absurd outcomes. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[3] Elinor Amit and Joshua D. Greene. You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science*, 23(8):861–868, 2012.

[4] Anonymous. Can large language models learn morality? a survey on definition, methodology and evaluation. *ACL ARR 2024 February Blind Submission*, February 2024. Paper Type: long, Research Area: Ethics, Bias, and Fairness, Contribution Types: Surveys, Languages Studied: English.

[5] Hannah Arendt. *Eichmann in Jerusalem: A Report on the Banality of Evil*. Viking Press, 1963.

[6] K. Carrie Armel, Aurelie Beaumel, and Antonio Rangel. Biasing simple choices by manipulating relative visual attention. *Judgment and Decision Making*, 3(5):396–403, Jun 2008.

[7] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[8] Roland Barthes. *Camera Lucida: Reflections on Photography*. Hill and Wang, 1981.

[9] Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshul Govil, Ponnurangam Kumaraguru, and Manas Gaur. Sage: Evaluating moral consistency in large language models, 2024.

[10] Alice Cai, Ian Arawjo, and Elena L. Glassman. Antagonistic ai, 2024.

[11] Michael Castelle. The linguistic ideologies of deep abusive language classification. In *Workshop on Abusive Language Online*, 2018.

[12] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Hate is not binary: Studying abusive behavior of gamergate on twitter, 2017.

[13] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning, 2022.

[14] Samuel Cooper and Sasha Lawson-Frost. Iris murdoch on moral vision. *Think*, 20(59):63–76, 2021.

[15] Damien L. Crone, Stefan Bode, Carsten Murawski, and Simon M. Laham. The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PLOS ONE*, January 2018.

[16] Jessica Dai. Beyond personhood: Agency, accountability, and the limits of anthropomorphic ethical analysis, 2024.

[17] Julian De Freitas and George A. Alvarez. Your visual system provides all the information you need to make moral judgments about generic visual events. *Cognition*, 178:133–146, 2018.

[18] Samuel Dooley, Rhea Sanjay Sukthanker, John P. Dickerson, Colin White, Frank Hutter, and Micah Goldblum. Rethinking bias mitigation: Fairer architectures make for fairer face recognition, 2023.

[19] John Drummond and Mark Timmons. Moral Phenomenology. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition, 2023.

[20] James Gordon Finlayson and Dafydd Huw Rees. Jürgen Habermas. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

[21] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms, 2021.

[22] Michel Foucault. Truth and power. In *Special Collection: CogNet*. The MIT Press, 2001.

[23] Minou Ghaffari and Susann Fiedler. The power of attention: Using eye gaze to predict other-regarding and moral choices. *Psychological Science*, 29(11):1878–1889, 2018. PMID: 30295569.

[24] Nelson Goodman. Seven strictures on similarity. In *Problems and Projects*. Bobs-Merril, 1972.

[25] Theodore Gracyk. Hume's Aesthetics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

[26] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press, Burlington, 2013. © 2013 Elsevier Inc.

[27] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[28] Donna J. Haraway. A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century. In *Simians, Cyborgs and Women: The Reinvention of Nature*, pages 149–181. Routledge, 1985.

[29] Donna J. Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599, 1988.

[30] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2023.

[31] R. W. Hepburn and Iris Murdoch. Symposium: Vision and choice in morality. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 30:14–58, 1956.

[32] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024.

[33] Yujin Jeong, Seongbeom Park, Suhong Moon, and Jinkyu Kim. Zero-shot visual commonsense immorality prediction, 2022.

[34] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. Can machines learn morality? the delphi experiment, 2022.

[35] Marianne Jørgensen and Louise Phillips. *Discourse Analysis as Theory and Method*. SAGE Publications, London; Thousand Oaks, CA; New Delhi, 2002.

[36] Rebecca M Kastner. Moral judgments and visual attention : An eye-tracking investigation. 2011.

[37] Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. Do moral judgment and reasoning capability of llms change with language? a study using the multilingual defining issues test, 2024.

[38] Douwe Kiela, Hamed Firooz, and Aravind Mohan. Hateful memes challenge and dataset for research on harmful multimodal content. `https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/`, May 2020. Meta.

[39] Karin Kreutzer. On the discursive construction of social entrepreneurship in pitch situations: The intertextual reproduction of business and social discourse by presenters and their audience. *Journal of Business Ethics*, 179:1071–1090, Jun 2022.

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[41] Johanna Lönngren. Exploring the discursive construction of ethics in an introductory engineering course. *Journal of Engineering Education*, 110(1):44–69, jan 2021.

[42] Rafal Maciag. Pretrained language models as containers of the discursive knowledge. *Comput. Sci. Math. Forum*, 8(1):93, 2023. Presented at the 2023 Summit of the International Society for the Study of Information (IS4SI 2023), Beijing, China, 14–16 August 2023.

[43] Marshall McLuhan. *Understanding Media: The Extensions of Man*. McGraw-Hill, 1964.

[44] Marshall McLuhan and Quentin Fiore. *The Medium is the Massage*. Bantam Books, 1967.

[45] Ben R. Newell and Mike E. Le Pelley. Perceptual but not complex moral judgments can be biased by exploiting the dynamics of eye-gaze. *Journal of Experimental Psychology: General*, 147(3):409 – 417, 2018.

[46] Chaim Noy. Moral discourse and argumentation in the public sphere: Museums and their visitors. *Discourse, Context Media*, 16:39–47, 2017.

[47] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[48] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

[49] Seongbeom Park, Suhong Moon, and Jinkyu Kim. Ensuring visual commonsense morality for text-to-image generation, 2023.

[50] Mark Pock, Andre Ye, and Jared Moore. Llms grasp morality in concept, 2023.

[51] Penny Powers. The philosophical foundations of foucaultian discourse analysis. *Critical Approaches to Discourse Analysis across Disciplines*, 1(2):18–34, 2007.

[52] Philip Pärnamets, Petter Johansson, Lars Hall, Christian Balkenius, Michael J. Spivey, and Daniel C. Richardson. Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, 112(13):4170–4175, 2015.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[54] Aida Ramezani and Yang Xu. Knowledge of cultural moral norms in large language models, 2023.

[55] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do, 2022.

[56] Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. Bert has a moral compass: Improvements of ethical and moral values of machines, 2019.

[57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[58] Gabriel Simmons. Moral mimicry: Large language models produce moral rationalizations tailored to political identity, 2023.

[59] Susan Sontag. *On Photography*. Farrar, Straus and Giroux, 1977.

[60] Susan Sontag. *Regarding the Pain of Others*. Picador, 2003.

[61] Susan Sontag. *On Photography*. RosettaBooks LLC, New York, first electronic edition edition, 2005. Published by arrangement with Farrar, Straus & Giroux.

[62] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947, March 2024.

[63] Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. A word on machine ethics: A response to jiang et al. (2021), 2021.

[64] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022.

[65] Andre Ye. And then the hammer broke: Reflections on machine ethics from feminist philosophy of science. In *Pacific University Philosophy Conference*, Forest Grove, Oregon, April 2024.

[66] Andre Ye, Jared Moore, Rose Novick, and Amy X. Zhang. Language models as critical thinking tools: A case study of philosophers, 2024.

[67] Yin Zhiyi. A longitudinal examination of foucault's theory of discourse. *Eurasian Journal of Applied Linguistics*, 9(1):152–160, February 2023.

[68] Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. Normbank: A knowledge bank of situational social norms, 2023.