Global Multidisciplinary Journal
eISSN: 2791-173X pISSN: 2791-2760

VOLUME04 ISSUE 09
Published 01-09-2025 *Page No. 21-29*

Unsupervised Feature Alignment: Ethical and Explainable Contrastive Approaches in Multimodal Artificial Intelligence Systems

Dr. Elias Thorne

Department of Computational Sciences, Meridian University

Dr. Sarah Vance
Institute for Ethical Artificial Intelligence

ABSTRACT

Background: The advent of Multimodal Artificial Intelligence has been accelerated by contrastive approaches to self-supervised learning, enabling systems to learn rich, robust feature representations without the need for expensive manual labeling. However, these "black box" models often produce high-dimensional latent spaces that are opaque to human interpretation, posing significant risks in high-stakes environments such as healthcare and criminal justice.

Methods: This study proposes a theoretical framework that bridges the gap between unsupervised contrastive learning and Explainable AI (XAI). We integrate principles of Granular Computing and Fuzzy Set Theory to impose interpretable structures upon the latent feature spaces generated by contrastive losses. Furthermore, we apply the National Institute of Standards and Technology (NIST) principles of explainability to evaluate the ethical standing of these systems.

Results: Our analysis demonstrates that while contrastive methods maximize feature richness, they often sacrifice semantic clarity. By applying granular modeling, we show that continuous feature vectors can be discretized into interpretable "information granules," thereby allowing for post-hoc explainability without retraining the foundational model. We further analyze the impact of confidence calibration on user trust.

Conclusions: We conclude that learning rich features without labels is viable for critical systems only when paired with robust XAI mechanisms. The integration of granular computing provides a mathematical foundation for extracting meaning from unlabeled data. We advocate for a "human-in-the-loop" governance model to ensure that contrastive AI systems remain ethical, transparent, and socially responsible.

KEYWORDS

Contrastive Learning, Explainable AI, Multimodal Systems, Granular Computing, AI Ethics, Unsupervised Learning, Feature Representation.

INTRODUCTION

The trajectory of modern Artificial Intelligence (AI) has been fundamentally altered by the emergence of Self-Supervised Learning (SSL). Historically, the success of deep learning architectures relied heavily on supervised learning paradigms, which require vast datasets annotated by human experts. While effective, this dependence on labeled data creates a significant bottleneck, particularly in specialized domains such as medical imaging or legal analytics where expert annotation is prohibitively expensive or scarce. Consequently, the field has shifted toward unsupervised and self-supervised approaches, specifically contrastive learning, which aims to learn rich features without labels by pulling semantically similar data points together in a latent space while pushing dissimilar points apart.

However, as these systems grow in complexity and capability, a paradox emerges: the very mechanisms that allow for the extraction of rich, nuanced features from unlabeled data often render the resulting models opaque. In the context of Multimodal AI—systems that process diverse data types such as text, images, and audio simultaneously—this opacity is compounded. The latent representations formed by contrastive losses are highly abstract, residing in high-dimensional spaces that defy simple interpretation. This "black box" nature creates a critical barrier to deployment in high-stakes decision-making environments. As Arrieta et al. note, the demand for Explainable Artificial Intelligence (XAI) is not merely technical but societal, driven by the need for responsible and transparent systems [1].

The tension between performance and explainability is arguably the defining challenge of the current AI epoch. While contrastive approaches yield state-of-the-art results in downstream tasks, their decision-making pathways remain obscure. This obscurity is problematic when viewed through the lens of regulatory frameworks and ethical standards. Phillips et al., in their work with the National Institute of Standards and Technology (NIST), outline four fundamental principles of explainable AI: explanation, meaningful, explanation accuracy, and knowledge limits [2]. Applying these principles to unsupervised systems, where the "ground truth" labels are absent during the training phase, presents unique difficulties. How does one explain a decision based on features that the model discovered autonomously, without human guidance?

This paper addresses this gap by proposing a framework for ethical and explainable contrastive learning. We argue that the richness of features learned without labels must be matched by a rigorous methodology for interpretation. To achieve this, we explore the intersection of contrastive learning, XAI, and Granular Computing—a paradigm described by Bargiela and Pedrycz as human-centric information processing [11]. By treating the learned features as "information granules," we can begin to bridge the semantic gap between mathematical vectors and human understanding.

The subsequent sections of this manuscript are organized as follows: Section 2 reviews the relevant literature regarding XAI taxonomies, the risks of post-hoc explainability, and the fundamentals of granular computing. Section 3 outlines our theoretical methodology for analyzing latent spaces. Section 4 presents a detailed analysis of feature richness versus interpretability. Section 5 discusses the specific implications for healthcare and ethics, expanding on the role of fuzzy modeling in clinical decision support. Finally, Section 6 offers concluding remarks on the future of transparent unsupervised learning.

2. RELATED WORK AND THEORETICAL BACKGROUND

To understand the challenge of interpreting features learned without labels, one must first survey the landscape of XAI and the ethical imperatives governing its use. The literature reveals a growing consensus that accuracy alone is insufficient for trust.

2.1 The Imperative of Explainability (XAI)

Explainable AI refers to the suite of techniques and methods that enable human users to comprehend and trust the results and output created by machine learning algorithms. Arrieta et al. provide a comprehensive taxonomy of XAI, distinguishing between transparent models (which are interpretable by design) and post-hoc explainability (techniques applied to opaque models after training) [1]. In the context of contrastive learning, which typically utilizes deep neural networks (e.g., ResNets or Transformers), the models fall squarely into the opaque category, necessitating robust post-hoc methods.

However, reliance on post-hoc explanation is not without risk. Vale, El-Sharif, and Ali highlight significant limitations in post-hoc methods, particularly regarding non-discrimination law [3]. They argue that approximations of a model's behavior may not faithfully represent the underlying decision logic, potentially masking bias or discriminatory patterns embedded in the unsupervised features. This is a critical observation for our study, as it suggests that "explaining" a contrastive model requires more than just a simplified surrogate; it requires a deep interrogation of the feature space itself.

2.2 Contrastive Learning and Multimodal Systems

Contrastive learning operates on the intuition of instance discrimination. By augmenting data (e.g., cropping an image or masking text) and training the network to recognize these augmented views as the same instance while distinguishing them from other instances, the model learns robust representations. These representations are "rich" because they capture high-level semantic structures necessary to distinguish between complex inputs. Yet, unlike supervised learning where features are optimized to predict a specific label (e.g., "cat" or "tumor"), contrastive features are optimized for geometric separation. This decoupling from explicit semantic labels makes the resulting features harder to map to human concepts.

2.3 Granular Computing and Fuzzy Logic

To bridge the gap between continuous latent vectors and discrete human concepts, we turn to Granular Computing. Zadeh introduced the concept of information granularity as a super-concept encompassing fuzzy sets, rough sets, and intervals [12]. The core idea is that human cognition operates on "granules"—clumps of similar objects or concepts—rather than precise numerical values. Keet describes granular computing as a paradigm for problem-solving that recognizes and exploits the granular nature of reality [13].

In the context of AI, Novak, Perfilieva, and Dvorak discuss fuzzy modeling as a method to handle vagueness [14]. This is particularly relevant for unlabeled features. A cluster of data points in a contrastive model's latent space might not correspond perfectly to a crisp category (like "malignant") but rather to a fuzzy set (like "suspicious texture"). Utilizing granular computing allows us to formalize these approximate relationships, turning mathematical proximity into linguistic descriptors.

2.4 Ethical Considerations

The deployment of these systems cannot be separated from their ethical context. Brendel et al. emphasize the need for ethical management of artificial intelligence, suggesting that organizations must adopt frameworks that prioritize transparency and accountability [15]. Furthermore, Shankheshwaria and Patel argue that building transparent models is essential for business applications, where trust determines adoption [16]. The literature suggests that as we move toward unsupervised learning, the burden of ethical assurance shifts from checking the labels (which don't exist) to auditing the learned representations for fairness and safety.

3. METHODOLOGY

This study employs a theoretical analysis and synthesis approach, integrating principles from machine learning, granular computing, and AI ethics to construct a framework for "Interpretable Contrastive Learning."

3.1 Theoretical Framework: The Granular Bridge

Our proposed methodology posits that the latent space generated by contrastive learning can be analyzed as a collection of information granules. We define a "semantic granule" within the high-dimensional space as a region where feature vectors exhibit high density and share commonalities that can be mapped to a human-understandable concept.

The process involves three stages:

1. Unsupervised Pre-training: A multimodal model (e.g., handling text and image) is trained using a contrastive loss function (such as InfoNCE) to minimize the distance between positive pairs and maximize the distance between negative pairs.

2. Granulation: We apply fuzzy clustering algorithms to the resulting feature space. Instead of forcing hard assignments (Cluster A vs. Cluster B), we assign membership degrees to various granules. This aligns with the work of Bargiela and Pedrycz [11], allowing for the modeling of ambiguity.

3. Semantic Mapping: Using a small set of "anchor" examples or external knowledge bases, we attempt to label these granules. For instance, a granule in a medical imaging model might be mapped to "irregular border," even if the model was never explicitly trained on border irregularity.

3.2 Evaluation Strategy

To evaluate the efficacy of this framework, we utilize the metrics proposed by Zhang, Liao, and Bellamy regarding the effect of explanation on trust calibration [5]. We analyze whether providing granular explanations (e.g., "The model grouped this image with 'irregular borders' with 0.8 membership") improves the user's ability to accept correct model predictions and reject incorrect ones. This distinguishes our approach from standard accuracy metrics; we are interested in the alignment between the model's internal state and the user's mental model.

4. RESULTS AND ANALYSIS

4.1 Feature Richness vs. Interpretability

The primary advantage of contrastive learning is the generation of rich features. In supervised learning, a model trained to classify "dog" vs. "cat" might ignore the background, lighting, or texture if those features do not help minimize the classification error. In contrast, contrastive models, which must distinguish a specific dog image from thousands of other images, tend to encode a vast amount of detail—color histograms, texture gradients, and contextual cues.

However, our analysis reveals that this richness is a double-edged sword for interpretability. The "Rashomon Effect" becomes prominent: there are multiple features that could explain the distinction between two data points. Without labels to constrain the model's focus, the model might distinguish two patients based on the calibration of the MRI machine rather than biological pathology.

Holzinger et al. note that explainable AI methods must facilitate a "human-in-the-loop" [10]. When we analyze the latent spaces of standard contrastive models, we find that the dimensions often do not correspond to disentangled concepts. Dimension 452 might encode a mix of "brightness" and "edge density." This entanglement renders simple linear probes insufficient for explanation.

4.2 The Efficacy of Granular Discretization

Applying the granular computing lens offers a solution. By analyzing the local topology of the feature space, we can identify "prototypical" examples that serve as the centers of granules. We find that fuzzy membership functions effectively capture the uncertainty inherent in unlabeled data.

For example, consider a multimodal system analyzing customer reviews (text) and product images. A standard clustering approach might force a product into "positive" or "negative" sentiment. A granular approach recognizes that a review might belong 0.6 to "functional satisfaction" and 0.4 to "aesthetic dissatisfaction." This nuance is preserved in the contrastive features but is often lost in downstream classification layers. Accessing the features directly through granular analysis recovers this information.

4.3 Trust Calibration

Zhang et al. demonstrated that confidence scores and explanations significantly impact trust calibration [5]. In our theoretical analysis of unsupervised systems, "confidence" is elusive because there is no ground truth probability distribution during pre-training. However, the density of the feature space can serve as a proxy. If a new data point falls into a sparse region of the latent space (a "gap" between granules), the system should report low confidence.

We find that "distance-to-prototype" explanations—common in contrastive learning—are only effective if the prototypes are meaningful. If the prototype for a decision is an adversarial example or an artifact, the explanation decreases trust. Therefore, the ethical deployment of these systems requires the rigorous curation of the "anchor points" used to define the semantic granules.

5. DISCUSSION AND EXTENDED ANALYSIS

The implications of learning rich features without labels extend far beyond technical architecture; they fundamentally reshape how we approach decision-making in critical sectors. To reach a holistic understanding of the subject, we must expand our scope to include specific applications in healthcare, the role of fuzzy modeling in managing uncertainty, and the governance structures required to maintain ethical integrity.

5.1 Healthcare Case Study: The Stakes of Unlabeled Learning

The healthcare sector represents the frontier where the promises and perils of AI collide most violently. Bhattacharya et al. pose the question of whether AI in healthcare is "hype, hope, or harm" [4]. In the context of contrastive learning, the "hope" is significant. Medical data is notoriously difficult to label; pathologists differ in their diagnoses, and annotating 3D volumetric scans is time-consuming. An AI system that can learn representations of disease pathology from millions of unlabeled scans could revolutionize diagnostics.

However, the "harm" potential is equally high. Amann et al. discuss the multidisciplinary perspective required for explainability in healthcare [7]. A contrastive model might cluster patients based on hidden covariates—such as the specific hospital where the scan was taken—rather than medical condition. This is known as "shortcut learning." In a supervised setting, we would see the model fail on a test set from a different hospital. In an unsupervised setting, this failure might go unnoticed until deployment because the model successfully minimized the contrastive loss by using the hospital tag as a discriminator.

Durán argues for a dissection of scientific explanation in AI (sXAI) specifically for medicine [8]. A medical explanation must be causal, not just correlational. Here, the limitation of contrastive approaches becomes stark. Contrastive learning identifies associations—data point A is similar to data point B. It does not identify causality—data point A has a tumor because of feature X.

To mitigate this, Antoniadi et al. highlight the opportunities for XAI in clinical decision support systems (CDSS)

[6]. We propose that CDSS utilizing unsupervised features must employ a "human-in-the-loop" verification step. Before a feature cluster is used for diagnosis, it must be validated by a clinician. For instance, if the model clusters a group of retinal scans together, an ophthalmologist must verify that the cluster corresponds to "diabetic retinopathy" and not "scans taken with Flash enabled." This transforms the AI from an autonomous diagnostician into a sophisticated pattern discovery tool.

5.2 Deep Dive: Granular Computing and Fuzzy Sets in Feature Extraction

To truly understand how we can make unlabeled features interpretable, we must delve deeper into the mathematical foundations of Granular Computing, as referenced by Zadeh, Keet, and Novak [12, 13, 14].

Standard machine learning often assumes that the world is made of crisp sets: an image is either a cat or not a cat. However, medical and sociological realities are rarely binary. A patient may be "somewhat hypertensive." A legal case may be "moderately precedent-setting." Contrastive learning, interesting enough, operates in a continuous space that supports this non-binary reality, but we frequently destroy this nuance when we force the output into a softmax classification.

Fuzzy Sets as Interpretability Layers:

Zadeh's theory of fuzzy sets [12] provides the logic for interpretation. In a fuzzy set, an element has a degree of membership ranging from 0 to 1. We can view the high-dimensional vector v generated by a contrastive model as existing within multiple fuzzy sets simultaneously.

Let us imagine a latent space for a dermatology AI. We can define fuzzy granules such as "Redness," "Asymmetry," and "Border Irregularity."

A specific image embedding x might have membership values:

$\mu_{Redness}(x) = 0.8$

$\mu_{Asymmetry}(x) = 0.2$

$\mu_{Irregularity}(x) = 0.6$

In a supervised system, we would need labels for redness, asymmetry, and irregularity to train these detectors. In our proposed unsupervised framework, we identify these granules via the statistical distribution of the data. We observe that a large cluster of data points shares a certain vector orientation. By examining the centroids of these clusters (the "prototypes" mentioned in section 4.2), a human expert can label the cluster as "Redness." Once labeled, the entire cluster becomes a "semantic granule."

Information Granularity and Abstraction:

Bargiela and Pedrycz emphasize that granular computing is about abstraction [11]. Humans handle complexity by ignoring details. When a doctor says "the patient is stable," they are using a high-level information granule that abstracts away thousands of biological metrics (heart rate variability, blood oxygenation, etc.).

Contrastive models "learn" these metrics. The challenge is to reverse-engineer the abstraction. By adjusting the "resolution" of our granulation (the size of the clusters we analyze), we can extract features at different levels of abstraction.

- Coarse Granularity: The model distinguishes "X-ray" vs "MRI".
- Fine Granularity: The model distinguishes "Viral Pneumonia" vs "Bacterial Pneumonia".

This hierarchical view of the latent space allows for multi-level explanations. A user can ask "Why is this similar?" and receive an answer at the appropriate level of granularity. This aligns with Cabitza et al.'s typology of explanation, which argues that the level of explanation must match the user's expertise [9].

5.3 Ethical Governance and Social Responsibility

The final, and perhaps most critical, pillar of our discussion concerns the governance of these systems. As AI permeates social infrastructure, the "black box" nature of unsupervised learning becomes a liability.

Algorithmic Bias in Unlabeled Data:

It is a misconception that removing labels removes bias. Camilleri notes that AI governance must address social responsibility [18]. Unlabeled data reflects the biases of the society that generated it. If a text corpus contains historical stereotypes, a contrastive language model (like a Large Language Model) will embed these stereotypes into the geometry of its feature space. "Doctor" will be geometrically closer to "Man," and "Nurse" to "Woman."

Because there are no labels to "correct" this during training, the bias is insidious. It is baked into the representation itself. Chi, Lurie, and Mulligan argue for reconfiguring diversity and inclusion frameworks for AI ethics [20]. In the context of contrastive learning, this means we must audit the data ingestion pipeline. We cannot rely on correcting the output; we must ensure the input distribution is diverse.

Ethical Assurance and The "Right to Explanation":

Burr and Leslie propose "Ethical Assurance" as a practical approach to responsible design [17]. This involves a continuous process of auditing and documentation. For unsupervised systems, we propose that "Ethical Assurance" requires the documentation of the latent space topology. Developers must prove that sensitive attributes (race, gender, age) are not the primary drivers of variance in the feature space.

This relates back to Vale et al.'s discussion on non-discrimination law [3]. If a bank uses an unsupervised model to segment customers for loan offers, and that model clusters people by zip code (a proxy for race), the bank is liable. The "rich features" learned by the model are, in this case, "toxic features."

Our proposed granular framework aids here. By identifying the granules that correlate with sensitive attributes, we can "prune" or "penalize" those dimensions, effectively surgically removing the bias from the representation without discarding the useful utility of the model.

Governance Protocols:

Char et al. identify ethical considerations specifically for ML in healthcare [19]. We can generalize this to a governance protocol for all critical unsupervised systems:

1. Data Provenance: Strict documentation of the source of unlabeled data.
2. Latent Space Auditing: Use of granular analysis to check for "shortcut learning" and bias.
3. Human-in-the-loop Labeling: Post-hoc labeling of feature clusters by diverse teams of experts.
4. Confidence Calibration: As discussed by Zhang et al. [5], systems must communicate their uncertainty. A low-density region in the latent space must trigger a "refer to human" protocol.

6. CONCLUSION

The evolution of Artificial Intelligence toward unsupervised and contrastive learning paradigms marks a significant leap in computational capability. The ability to learn rich, robust features without the constraints of manual labeling unlocks the potential of vast, unstructured multimodal datasets. However, this power comes at the cost of transparency. The high-dimensional geometric relationships that define these "rich features" are inherently difficult for human cognition to grasp.

This paper has argued that the deployment of such systems in critical domains—healthcare, law, and finance—is ethically impermissible without a concurrent framework for explainability. We have proposed that Granular Computing and Fuzzy Set Theory offer the mathematical bridges necessary to span the chasm between continuous vector spaces and discrete human concepts. By treating features as information granules, we can impose semantic structure on the "black box."

Furthermore, we have highlighted that accuracy and richness are not the only metrics of success. Trust, fairness, and non-discrimination are equally vital. The literature confirms that post-hoc explanations have limitations, and blindly trusting unsupervised clusters can perpetuate societal biases. Therefore, we conclude that the future of Multimodal AI lies not just in better contrastive loss functions, but in "Ethical Contrastive Systems"—architectures that are designed from the ground up to be audited, interpreted, and governed. Only by illuminating the latent space can we ensure that the features we learn are not just rich, but also right.

REFERENCES

1. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 2020, 58, 82–115.

2. Phillips, P.J.; Hahn, C.A.; Fontana, P.C.; Broniatowski, D.A.; Przybocki, M.A. Four Principles of Explainable Artificial Intelligence; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020; Volume 18.

3. Vale, D.; El-Sharif, A.; Ali, M. Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. AI Ethics 2022, 2, 815–826.

4. Bhattacharya, S.; Pradhan, K.B.; Bashar, M.A.; Tripathi, S.; Semwal, J.; Marzo, R.R.; Bhattacharya, S.; Singh, A. Artificial intelligence enabled healthcare: A hype, hope or harm. J. Fam. Med. Prim. Care 2019, 8, 3461–3464.

5. Zhang, Y.; Liao, Q.V.; Bellamy, R.K.E. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 295–305.

6. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. Appl. Sci. 2021, 11, 5088.

7. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I.; the Precise, Q.c. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Med. Inform. Decis. Mak. 2020, 20, 310.

8. Durán, J.M. Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. Artif. Intell. 2021, 297, 103498.

9. Cabitza, F.; Campagner, A.; Malgieri, G.; Natali, C.; Schneeberger, D.; Stoeger, K.; Holzinger, A. Quod erat demonstrandum?—Towards a typology of the concept of explanation for the design of explainable AI. Expert Syst. Appl. 2023, 213, 118888.

10. Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; Samek, W. Explainable AI methods—A brief overview. In Proceedings of the xxAI—Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, Vienna, Austria, 12–18 July 2020; Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., Samek, W., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 13–38.

11. Bargiela, A.; Pedrycz, W. Human-Centric Information Processing through Granular Modelling; Springer Science & Business Media: Dordrecht, The Netherlands, 2009; Volume 182.

12. Zadeh, L.A. Fuzzy sets and information granularity. In Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers; World Scientific: Singapore, 1979; pp. 433–448.

13. Keet, C.M. Granular computing. In Encyclopedia of Systems Biology; Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H., Eds.; Springer: New York, NY, USA, 2013; p. 849.

14. Novák, V.; Perfilieva, I.; Dvoˇrák, A. What is fuzzy modeling. In Insight into Fuzzy Modeling; John Wiley & Sons: Hoboken, NJ, USA, 2016; pp. 3–10.

15. Brendel, A.B., Mirbabaie, M., Lembcke, T.B. and Hofeditz, L., 2021. Ethical management of artificial intelligence. Sustainability, 13(4), p.1974.

16. Yashika Vipulbhai Shankheshwaria, & Dip Bharatbhai Patel. (2025). Explainable AI in Machine Learning: Building Transparent Models for Business Applications. Frontiers in Emerging Artificial Intelligence and Machine Learning, 2(08), 08–15.

17. Burr, C. and Leslie, D., 2023. Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. AI and Ethics, 3(1), pp.73-98.

18. Camilleri, M.A., 2023. Artificial intelligence governance: Ethical considerations and implications for social responsibility. Expert Systems, p.e13406.

19. Char, D.S., Abràmoff, M.D. and Feudtner, C., 2020. Identifying ethical considerations for machine learning healthcare applications. The American Journal of Bioethics, 20(11), pp.7-17.

20. Chi, N., Lurie, E. and Mulligan, D.K., 2021, July. Reconfiguring diversity and inclusion for AI ethics. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 447-457).