Qeios

Peer Review

Review of: "Undefinable True Target Learning: Towards Learning with Democratic Supervision"

Renato Vicente[1]

1. Applied Mathematics, Universidade de São Paulo, Brazil

General Appraisal

I think the premise of questioning the objectivity of ground truth is timely and resonates with contemporary discourse in the machine learning community, particularly within the subfields of data perspectivism, subjective NLP, and aleatoric uncertainty quantification in medical imaging.

However, the manuscript suffers from substantial deficiencies that undermine its contribution. Primarily, the claim of novelty regarding the "Non-Existence Assumption" is overstated due to a lack of engagement with established literature on Perspectivist Machine Learning and Crowd Truth, which have explored this exact premise for over a decade. Furthermore, the terminological choice of "Democratic Supervision" to describe a human-AI feedback loop is semantically inconsistent with established definitions in Political Science and social computing, creating conceptual confusion. Mathematically, the manuscript lacks formal proofs and relies on tautological hypotheses. The "Practical Solutions" section relies heavily on self-citation of the author's prior work on One-Step Abductive Multi-Target Learning (OSAMTL) without providing new comparative empirical evidence against state-of-the-art baselines.

Relation with Previous Literature

The manuscript's central claim to originality lies in "explicitly positing that the TT does not exist in the real world". The author contrasts this position with Supervised Learning, Weakly Supervised Learning, and Reinforcement Learning, arguing that existing paradigms largely assume an objective TT exists, even if observed inaccurately. The author asserts that "few works explicitly reject the premise that the TT objectively exists in the real world".

doi.org/10.32388/7DY3QX

This assertion is factually incorrect and ignores a substantial body of literature that has explicitly rejected the "single truth" assumption for many years. To understand the magnitude of this oversight, it is necessary to map the existing landscape of Data Perspectivism and Subjective Ground Truth, which frames the author's contribution not as a novel discovery, but as a late entrant to an established debate.

For example, the work of Aroyo and Welty, particularly "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation" (2015), directly contradicts the manuscript's claim that current paradigms assume existence. They argue that disagreement among annotators is not noise to be eliminated but a valuable signal that captures the inherent ambiguity of the input data. They propose Crowd Truth metrics that model the vector of disagreement rather than collapsing it to a single label. This is conceptually identical to the author's premise of "Undefinable True Target," yet this work is not referenced in the manuscript.

Also "Learning from Disagreement: A Survey" by Uma et al. (2021) provides a comprehensive taxonomy of methods that operate without a single ground truth. This survey details approaches using Soft Labels (probability distributions over classes) and Multi-Task Learning to model individual annotators, covering the exact theoretical ground the author attempts to claim as novel.

Cabitza et al. (2023) formalized this movement by defining Weak Perspectivism (where data is aggregated, but metadata about disagreement is preserved) versus Strong Perspectivism (where models are trained to predict the distribution of opinions or individual perspectives, explicitly rejecting the existence of a single true label). The manuscript's failure to engage with the distinction between Weak and Strong Perspectivism renders its theoretical contribution somewhat redundant. The "Non-Existence Assumption" presented in Section 3 of the manuscript is effectively a restatement of Strong Perspectivism without the accompanying rigor or reference to the community that defined it.

The manuscript relies heavily on examples from medical imaging, specifically Helicobacter pylori segmentation and breast cancer tumor segmentation. The author argues that in these domains, the True Target is undefinable. While it is true that inter-rater variability is high in histopathology, the medical imaging community has addressed the "non-existence" of a perfect ground truth for decades through probabilistic frameworks.

The STAPLE (Simultaneous Truth and Performance Level Estimation) algorithm, developed by Warfield et al. (2004), is the standard method in medical imaging for handling missing ground truth. STAPLE computes a probabilistic estimate of the true segmentation by weighting raters based on their estimated performance, explicitly acknowledging that no single rater provides the "True Target." The manuscript's

failure to discuss STAPLE or its modern deep learning variants (e.g., Deep STAPLE) creates a significant gap in the literature review.

Moreover, recent work on Soft Segmentation (e.g., SoftSeg) and Aleatoric Uncertainty Quantification deals explicitly with the "undefinable" nature of boundaries. In these frameworks, the model predicts a variance or entropy map alongside the segmentation, quantifying the irreducible ambiguity of the data. This is a direct implementation of the “Non-Existence Assumption” where the model admits it cannot know the boundary precisely because the boundary is inherently fuzzy. By ignoring Aleatoric Uncertainty, the manuscript overlooks the current state-of-the-art solution to the problem it poses.

The manuscript also effectively reinvents the wheel of Strong Perspectivism but labels it "UTTL." I would suggest to the author pivoting from claiming the discovery of this problem to proposing a specific, novel method for solving it that outperforms existing Perspectivist or Uncertainty-aware methods. I think, in its current form, the theoretical contribution is diluted by the lack of acknowledgment of this foundational literature.

In summary, the manuscript identifies a valid and critical problem: the reliance on illusory ground truth. However, by claiming this insight is novel and explicitly positing the non-existence of TT as a "radical stance," the author ignores the massive intellectual progress made in Subjective NLP and Uncertainty Quantification.

Terminology

A significant portion of the manuscript is dedicated to the concept of Learning with Democratic Supervision (LDS). The author defines LDS as a scenario where "both domain and AI experts collaboratively contribute to data preparation, forming a more balanced and participatory (democratic) framework." I believe this definition is problematic on semantic, theoretical, and practical levels.

The term "Democracy" carries heavy semantic weight, implying a governance structure involving a plurality of voices (*demos*), rights, voting mechanisms, and institutional oversight. In the context of technology and AI, “Democratic Supervision” typically refers to oversight by citizens, elected bodies, or civil society over algorithms or intelligence agencies to ensure they align with public values. “Democratizing AI” refers to lowering the barrier to entry so that non-experts can create or use AI tools. “Crowdsourcing” involves large groups of diverse human annotators providing inputs, often aggregated via majority vote or consensus.

I think the author's usage of "Democratic" to describe a feedback loop between a single Domain Expert and a single AI model is a category error. An AI model is not a constituent; it does not have rights, perspectives, or "votes" in the political sense. It is a tool. Describing the interaction between a user and a tool as "democratic" anthropomorphizes the software and obscures the power dynamics. The Domain Expert (human) ultimately accepts or rejects the AI's suggestions. If the AI "votes" against the human, it is based on its objective function, not a "perspective."

The literature refers to this setup as Human-in-the-Loop, Interactive Machine Learning, or Collaborative Learning. Using "Democratic" adds unnecessary confusion and conflicts with the established usage in AI Ethics and Algorithmic Governance, where "Democratic Supervision" is a regulatory concept, not a training loop.

The manuscript also fails to compare LDS with Jury Learning (Gordon et al., 2022), a paradigm that genuinely embodies democratic principles in AI. In Jury Learning, the model output is determined by a "jury" of diverse human annotators who deliberate or vote. This approach explicitly models the demographics and perspectives of the jurors to reach a verdict that reflects community standards.

In contrast, UTTL's LDS is a closed loop between one human and one machine. By ignoring Jury Learning, the author misses the opportunity to position his work against a framework that successfully implements the democratic metaphor. The LDS framework, compared to Jury Learning, lacks the diversity of human perspectives that makes democratic processes robust.

The Concept of "AI Expert"

The manuscript elevates the algorithm to the status of an "AI Expert" (AIE) that "collaborates" with the Domain Expert (DE). I think that is problematic.

If the AIE is trained solely on the data provided by the DE, it cannot be an independent "expert." It can only reflect the statistical correlations found in the DE's data. If the DE's data is biased or flawed, the AIE will mimic those flaws. It cannot offer a "second opinion" in a democratic sense because it is dependent on the first opinion. If the supervision is "Democratic" in the sense that the AIE's output is fused with the DE's input to retrain the model, this creates a feedback loop. The model confirms its own biases, potentially leading to confirmation bias rather than a true discovery of "undefinable" truth. Also, the manuscript mentions an "accumulated knowledge set K" but treats it abstractly. In the author's prior work on OSAMTL, this K represented logic rules. If the "AI Expert" is actually a Symbolic Logic Module

(neuro-symbolic AI), this should be explicitly stated. A logic module enforces consistency, which is different from "democratic" voting. Consistency checking is a constraint, not a *vote*.

Theorems

The author presents four theorems that require some attention.

Theorem 1 seems to contradict the manuscript's core premise. The paper argues that "TT does not exist." If TT does not exist, the set $prop(TT)$ is undefined or empty. One cannot define a subset relation $\subseteq$ against a non-existent set. This is a fundamental logical inconsistency. If the author means "Latent Ideal Target," they must define it as such, but they explicitly rejected the "Existence Assumption."

Also, the proof relies on the hypothesis that "Domain expert knowledge is correct." Thus, Theorem 1 essentially says: "If we assume the expert is correct, then the expert is correct." But, this is a tautology.

Theorem 3 relies on the Condorcet Jury Theorem logic, namely, that aggregating diverse imperfect voters yields a better result. However, Condorcet's theorem requires that errors be independent. In the UTTL setup, the AIE is likely trained on data similar to the DE's, or refined from it. Therefore, the errors of the DE and AIE are highly correlated. Without proving the independence of the AIE and DE, Theorem 3 does not hold. Fusing two correlated sources (e.g., a teacher and a student trained by the teacher) does not necessarily increase information; it often reinforces shared errors. Unfortunately, the manuscript offers no mathematical proof of independence.

Practical Solutions

The "Practical Solutions" section lacks comparative experimental results against relevant baselines. To claim that UTTL is a superior paradigm, the author must compare it against:

1. Standard Aggregation: Majority Voting, Dawid-Skene.
2. Crowd Layer Methods: (Rodrigues & Pereira, 2018) , which explicitly model annotator matrices.
3. Uncertainty-Aware Methods: Bayesian Ensembles or Monte Carlo Dropout that quantify aleatoric uncertainty.
4. Soft-Label Learning: Training directly on the raw distribution of annotator votes.

Without these comparisons, the manuscript fails to prove that "Democratic Fusion" provides any performance benefit over standard soft-label training or uncertainty quantification. The claim that UTTL

"exemplifies the essence of LDS" remains an unproven assertion.

Recommendations

I think the manuscript identifies a critical philosophical and practical bottleneck in machine learning: the reliance on objective ground truth in an inherently subjective world. The author's intuition that we must move from finding a hidden truth to constructing a robust consensus through collaboration is sound and aligns with the frontier of Data Perspectivism and Human-Centric AI.

However, the current execution is flawed by a lack of engagement with prior literature, confusing terminology, and insufficient rigor. By framing established ideas as radical novelties and using the loaded term "Democracy" for a human-machine loop, the manuscript obscures its potential contribution. The "AI Expert" seems to me more like a neuro-symbolic logic module, and the "Democratic Supervision" is collaborative refinement.

If the author reframes the work as a Neuro-Symbolic Framework for Learning from Subjective Data, rigorously defines the "Knowledge Base" role, and benchmarks against modern Uncertainty Quantification and Crowd Truth methods, this research could offer a significant contribution to the field of Weakly Supervised Learning.

In its current form, however, I believe it does not yet fully meet the standards of rigor and novelty expected for a significant contribution.

Declarations

Potential competing interests: No potential competing interests to declare.