Human-centric annotation of multi-modal data: A framework perspective

Clément BELIVEAU[1, 2, 3, 4], Yann-Romain KECHABIA[4],
Cyril RAY [4], Maeve PETIT, Fabienne LAJONCHERE [3] , John PUENTES [1, 2]

[1]IMT Atlantique, Brest, FRANCE
[2]Lab-STICC CNRS UMR 6285, Brest, FRANCE
[3]Naval Group, CEMIS, Ollioules, FRANCE
[4]Ecole navale, IRENav, Lanveoc, FRANCE

clement.beliveau@naval-group.com

Abstract

Data annotation is a cognitively rich process shaped by perception, judgment, and variability. However, it is traditionally viewed as a mechanical and time-consuming task for humans, and is therefore often automated or outsourced. This paper considers annotation as a Human-Centric cognitive process and proposes a generic framework in which annotation supports humans in interpreting complex multimodal data and operational information. Through a taxonomy of human-in-the-loop semi-automatic pipelines, we show how AI-generated annotations can interact with humans. We apply this framework to a naval defense use case, specifically maritime surveillance activity. In this high-stakes environment, where operators must process continuous and heterogeneous information streams, annotation acts as a cognitive aid. It can ultimately become a decision-support tool that reduces cognitive load and preserves situational awareness, rather than remaining a simple data-labeling mechanism.

Keywords

Multi-modal data, Annotation, Data-Centric Artificial Intelligence, Human-Centric design, Human-In-The-Loop, Decision support.

1 Introduction

From finance to healthcare, from logistics to advertising, the modern world is awash in data, fueling the rapid advancement and integration of artificial intelligence (AI) across nearly every sector [25]. In this data saturated world, annotation is a foundational step in any machine learning pipeline, but from a cognitive ergonomics perspective it could also be seen as a form of cognitive externalization processes and serve as an efficient tool to help humans when dealing with complex and heterogeneous information [13, 22]. While often reduced to a mechanical labeling task, annotation is in reality shaped by human perception, judgment, and variability. In a Data-Centric AI paradigm [36], the quality of annotated data directly conditions its usage, making annotation an essential lever for system reliability. However, the human dimension of this process remains largely overlooked: who annotates, how, and under what cognitive conditions, are questions rarely addressed in the literature. Furthermore, as AI systems are increasingly deployed in operational settings, the boundary between annotation as a machine learning training step and annotation as a real-time decision-support tool becomes increasingly blurred. This raises a fundamental question: can annotation be reframed as a cognitive aid for humans confronted with complex, multi-modal, and ambiguous information, instead of merely as a resource for machine learning?

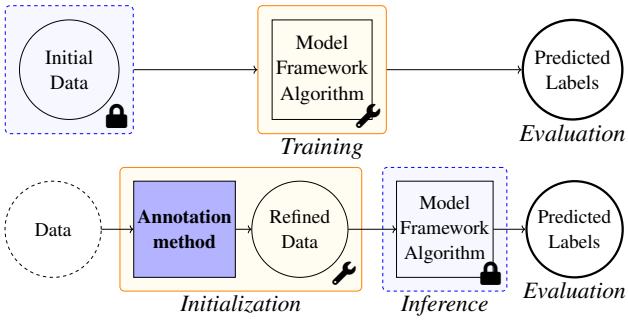

Figure 1: Model-Centric *(top)* and Data-centric *(bottom)* machine learning pipelines, 🔒 = frozen block – 🔧 = trainable block

Designed as a cognitive aid, data annotation requires to be implemented within a comprehensive, dynamic, and iterative improvement process, in which contextual nuances are swiftly captured. Regarding data annotation pipelines, currently two main paradigms prevail in AI: *Model-Centric* versus *Data-Centric* (Fig. 1). The Model-Centric one assumes that annotation is secondary, annotated data are available, data quality is taken for granted, and prioritizes algorithmic improvements through training techniques or architecture design. The Data-Centric paradigm shifts the focus to the quality, structure, and relevance of the data it-

self as the primary lever for enhancing model performance, to refine predictions by improving the model input [36].
Once a pipeline is trained and tuned with either paradigm, it is frozen and used for inference on new data (Fig. 2). This leads to the operator being a passive component of the pipeline. However, some critical sectors like defense, healthcare, or finance, cannot entirely remove humans from the decision process. For instance, in the defense domain, military operations increasingly rely on vast networks of sensors, technologies, and AI-based systems. Yet paradoxically, while defense operators are flooded with sensors data, they are often reluctant to share authority with autonomous or semi-autonomous decision-making systems. This tension raises a critical question: ***In high-stakes environments where human oversight is essential and data are overwhelming, what role can annotation play?***

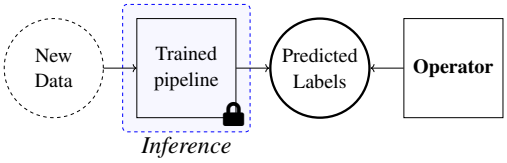

Figure 2: Inference phase where the operator uses the predicted labels

Hence, in this paper, we draw inspiration from the Data-Centric approach integrated with a Human-Centric paradigm, emphasizing the foundational yet often undervalued task of data **annotation** that precedes model training.
Our work proposes a Human-Centered approach (Fig. 3) where annotation serves as a decision-support mechanism for humans, who remain central system components in the interpretation of complex multi-modal data and information. The proposed methodology is illustrated through a naval defense use-case in Section 4. More specifically it is applied to the activity of maritime surveillance deployed in semaphore stations and frigates. Rather than emphasizing system implementation, this paper adopts a conceptual perspective and investigates the interaction between human operators and an annotation framework, laying the groundwork for future instantiations and operational deployment.

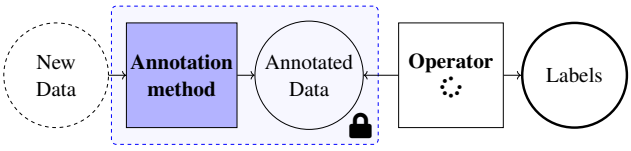

Figure 3: Position of the operator within a Human-Centered annotation paradigm

The article is organized as follows: Section 2 introduces a taxonomy of human-in-the-loop pipelines. The concept of data annotation, its use in the literature, and how it could be framed as a tool in a Human-Centered paradigm are presented in Section 3. The resulting concept is applied to the defense domain in Section 4, focusing on naval operators and instantiates the proposed multi-modal data annotation framework. Main resulting insights are discussed in section 5. Conclusions and perspectives are outlined in Section 6.

2 Human-in-the-loop for annotation

Annotation materializes the results from a succession of decisions and choices, aimed at summarizing insights on human perception at a given time, in order to train machine learning models [45]. Annotation can be fully manual, as with Amazon mechanical Turks [5] or crowd-sourcing [11, 27]. It can also be automatic, leveraging AI models trained on already annotated data [19, 37]. Finally, semi automatic ones combine human and AI to accomplish annotation [43]. Our framework moves away from a Model-Centric paradigm in which humans passively annotate datasets for machine learning. Instead, it promotes collaboration between humans and AI, where annotation becomes a tool to support human understanding. This approach aligns with hybrid intelligence research, which emphasizes the complementary roles of humans and AI in complex decision-making environments [10].
This section introduces the typology of Human-In-The-Loop (HITL) strategies for semi-automatic data annotation [43]. The goal is to understand how human and AI can interact, collaborate, and annotate together, and therefore, allow choosing the best HITL strategies for each use case and operator. Interactions between humans and several AI agents have not been previously studied in this particular context. We respectively describe as "Human" a given set of human operators and as "AI" a given set of models, within a simplified annotation process. We subdivide these strategies into five main types described below. Each one is illustrated by a schematic diagram. The legend of symbols is shown in Table 1.

Table 1: Legend and Symbols for annotation schemes

Legend	Description
H	Human
AI	Artificial Intelligence
A, A'	First and intermediate annotations
F	Final annotation

Symbols

◇	Annotation	(D)	Raw data
□	Annotation agent	⊕	Fusion component

2.1 Edge-Case

This pipeline is similar to a fully automatic one; the AI model processes the entire dataset, and the human acts as a judge. Instances where the model's confidence falls below a predefined threshold are flagged for review. A human annotator then inspects and corrects these *edge-cases*, which often lie near the model's decision boundary. This approach reduces cognitive load by filtering out trivial cases and concentrates human effort on the most uncertain or ambiguous

instances [27, 29]. However, AI model's errors in trivial examples might remain undetected [1], introducing biases in the annotated dataset (Fig. 4).

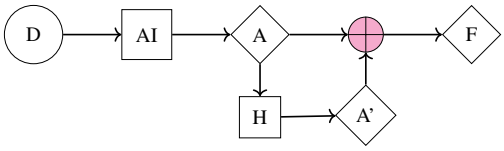

Figure 4: Block diagram of a semi-automatic edge-case pipeline (the human checks only automatic annotations that were flagged; automatic and human annotations are then fused)

2.2 Suggestion

The *suggestion* pipeline is conceptually closer to the manual approach; humans remain the primary annotators, while the AI can be summoned to suggest an annotation of a given instance. Operators can accept, modify, or reject these suggestions based on their judgment. This setup preserves human oversight and mitigates over-reliance on the model, leveraging AI assistance to ameliorate efficiency and consistency (Fig. 5).

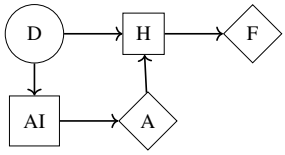

Figure 5: Block diagram of a semi-automatic suggestion pipeline (the final annotation is human - based on AI suggestions)

2.3 Iterative

In the *iterative refinement* pipeline, annotation is performed through alternating passes between the AI model and the human annotator. After each iteration, the model can be updated based on corrections provided by the human (active learning or reinforcement learning) [28] [23] [21]. This process continues until a predefined quality threshold is met [34]. The approach maintains expert involvement throughout the annotation, introducing the risk of human over-reliance on the model's suggestions, potentially limiting critical review (Fig. 6).

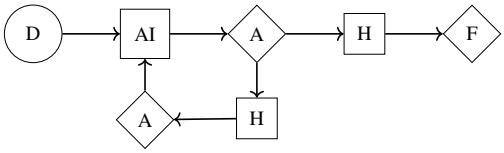

Figure 6: Block diagram of a semi-automatic iterative refinement pipeline (the final annotation is the result of a succession of human and AI annotations)

2.4 Few-shot

In this setting, a small part of the data is firstly annotated by humans, usually experts. These examples are then used to fine-tune or prompt in a *few-shot* setting to the model, which subsequently annotates the remaining data automatically. Additional human annotations may be requested if the model outputs fail to meet predefined quality thresholds. While this approach significantly reduces manual annotation efforts [15, 30], it remains susceptible to model severe inconsistencies and propagation of errors (Fig. 7).

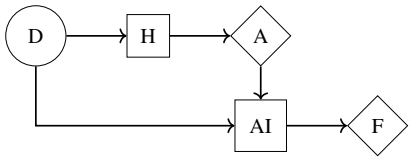

Figure 7: Block diagram of a semi-automatic few-shot pipeline (human annotations are used to automatically generate the final annotation)

2.5 Challenger

In the *challenger* pipeline, human and AI annotations are independently generated and compared to obtain the final annotation. Discrepancies trigger a dispute-resolution step, which may involve human review or additional decision logic. This design aims to mitigate human over-reliance on AI requiring the annotator to engage continually with each instance, while still benefiting from automated error detection and redundancy (Fig. 8).

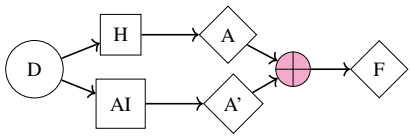

Figure 8: Block diagram of a semi-automatic challenger pipeline (human and AI annotate in parallel - the annotations are then fused)

3 Annotation properties and process

Annotation can be defined as a process through which humans add context, dimension, or information to existing data [19]. This section explores both its fundamental properties and the cognitive processes to derive annotations from raw data. On this basis, we introduce a generic framework in which annotation serves as a cognitive aid to support human interpretation and decision-making.

3.1 Fundamental properties

Annotation takes many forms depending of the context and goal, ranging from graphical to written representations. Initially formalized in textual contexts as "a note added to a text" [19], nowadays, annotation is everywhere in our dailylife:

- **Shopping mall**, a red dot "you are here" on a map or a colored arrow on a floor are graphical ones.
- **Television**, subtitles translating a movie or the crawling ticker during the news are written ones.
- **Restaurant**, reviews from previous customers or the number of stars they have assigned are both.

Varying from informal notes to formally structured labels, annotation is used to enrich data by adding contextual, structural, or semantic information, thus facilitating its understanding [16]. Annotation is intended for someone or something, a user or an AI model. Informal annotations can support human interpretation when exploring new data sources [35], whereas formal annotations typically consist of normalized labels that encode well-defined concepts. We identify four primary types of data that can be annotated: **textual, visual, numerical, and acoustic**, representing fundamental modalities to which any real-world data can be reduced.

3.1.1 Constructed from human perception

Annotation is based on how humans perceive input data. In this regard Gestalt principles and Cognitive Vision Theory (CVT) provide complementary perspectives on how humans perceive visual inputs. Gestalt principles explain how the human visual system organizes elements into coherent structures based on proximity, similarity, continuity, closure, and common fate [40]. CVT formalizes the layer of interpretation through which humans process and understand information [20] [44]. In spite of being initially formulated for visual stimuli, such concepts could transfer across various types of data [24].
Indeed, during annotation, humans establish a semantic interpretation of a situation that involves three levels [8]:

- **Low level**, corresponds to the perception of raw sensory features.
- **Mid level**, involves structural analysis and logical inference.
- **High level**, integrates abstract reasoning and prior knowledge.

Interpretation levels influence the cognitive effort during annotation. Specifically, high-level reasoning is typically an active, attention-demanding process [32], whereas low- and mid-level perception often occur unconsciously, governed by innate perceptual mechanisms [4]. In operational environments such as maritime surveillance, operators must rapidly shift between these levels to integrate heterogeneous signals and maintain a dynamic representation of the situation, reflecting situation awareness mechanisms that support perception, comprehension, and projection in complex, evolving environments [13, 42]. Understanding these perceptual mechanisms provides a foundation for designing annotation tools and workflows that optimize cognitive load, allowing operators to focus on high-value tasks.

3.1.2 Abstraction levels

According to examined works, we propose a classification of annotations into four hierarchical layers, each corresponding to a different level of abstraction:

1. **Surface layer**, refers to annotations based on raw, low-level features of the data. It is rarely treated as a distinct annotation layer in the literature [21].
2. **Structural layer**, captures the internal organization and structure of the data. Through global features, it identifies units and their spatial or temporal arrangement, often without interpreting their semantic content [26] [39].
3. **Semantic layer**, encodes standardized and widely understood meanings using shared conceptual frameworks. These annotations, based on common concepts, aim for consistency across systems and annotators [1, 20, 23].
4. **Interpretative layer**, represents subjective, high-level data interpretations, involving emotional, social, or contextual judgment. This annotation layer is somewhat recent, and unfolded initially from sentiment and social science fields, where human perception and affective response play a central role [45].

3.2 Annotation as, a human process

The role humans play in the annotation process is an essential topic rarely discussed, which could help broaden AI capabilities by grasping how human intelligence works in this particular case [6].

3.2.1 Human strategies

The cognitive process deployed by humans to realize manual annotations is called a strategy. It represents the user's plan of action, formed by chains of tactics and moves. Annotation tactics are the steps, actions, and choices deployed by humans to move forward in order to carry out an annotation. Tactics are formed by annotation moves that are basic actions of thought performed by the annotator (zoom-in, zoom-out, compare, measure, define, etc.) [8]. An annotation strategy might include, for instance, a fine-tuning tactic, where the operator refines or adjusts an annotation to achieve greater accuracy afterward [2].
Considering these strategies when designing annotation tools might be useful to encourage or avoid certain tactics. Although, even if strategies vary between annotators, trends emerge. For instance, individuals are likely to prefer starting with less accurate tactics involving perception, rather than cognitively demanding ones. Also, individuals applying less accurate methods tend to be willing to continue refining their work [8, 34].

3.2.2 Human biases

Humans are inherently sensitive to cognitive, perceptual, and emotional biases, which can affect annotation quality and may be exacerbated under conditions of cognitive

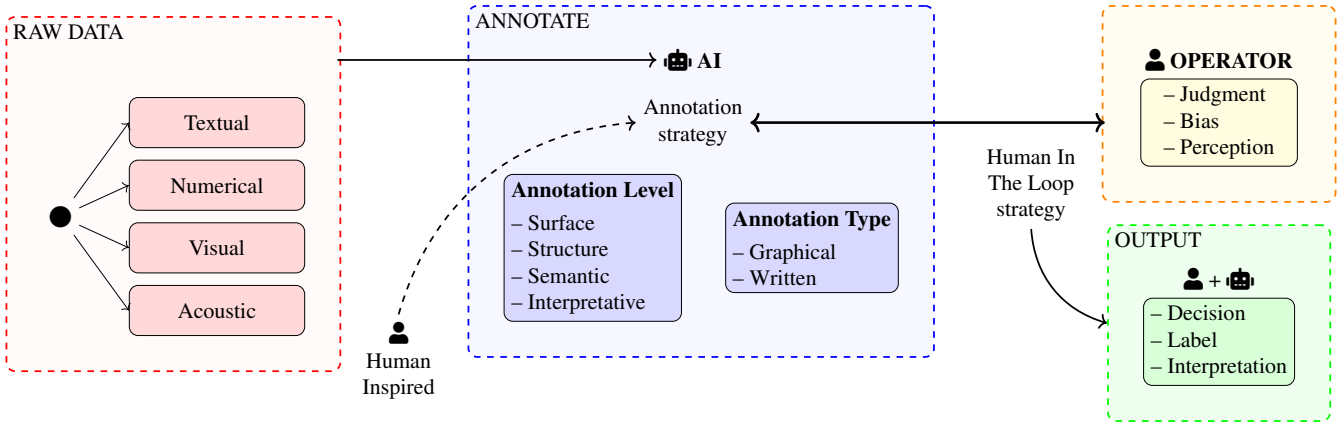

Figure 9: Proposed generic Human-Centered annotation framework for multi-modal data

overload [38]. Some biases, such as distrust or overconfidence, arise from the nature of the task itself, while others emerge from the structure and sequencing of the annotation process. Similar to surveys, annotation campaigns involve repeated human judgments under controlled conditions; in such settings, the order in which tasks are presented can induce contrast and assimilation effects [2].

The data being annotated can also generate variability in perception among annotators, potentially leading to confusion between certain labels [6]. This challenge is amplified by the heterogeneous and multi-modal nature of datasets, which often combine simple instances with complex ones of different modalities. Consequently, the difficulty of annotation can fluctuate dramatically, ranging from trivial to expert-level depending on the label schema. To mitigate such issues, adaptive labeling interfaces that dynamically adjust to the annotator's perceived task difficulty have been proposed [7].

How annotators perceive information is shaped by their prior experience and domain knowledge [11]. Individual's bias, expertise, or fatigue can influence the labels they assign [2, 9], with potential consequences for downstream model performance. Annotation is therefore not a purely technical task but a cognitively rich process shaped by perception, judgment, and variability.

By integrating all these components, we propose a generic Human-Centered annotation framework (Fig. 9), that can be tailored to specific use cases depending on the available data and the desired human-in-the-loop (HITL) configuration, as illustrated in Section 4 for a defense application. Especially suited for multi-modal data, this flexible framework searches to strengthen human decision-making, interpretation, and labeling capabilities. Based on contextual information an AI agent generates an annotation strategy designed to reflect human reasoning processes. The resulting annotation can thus become a cognitive artifact supporting operators' sense-making processes in complex environments, helping them externalize intermediate interpretations and stabilize evolving hypotheses during decision making [10, 22].

4 Application to a defense use case

Leveraging annotation not merely as a technical aid but as a cognitive support for humans remains largely unexplored both in hybrid-intelligence literature and in defense contexts where it may play a key role in supporting human decision-making in complex operational environments. This insight was reinforced by field observations conducted to model maritime surveillance activities. The pipeline of the current activity, presented in Fig. 10, is the result of a series of in situ observations and semi-structured interviews realized with a frigate crew and ten operators across four semaphores.

4.1 Multiplication of signals

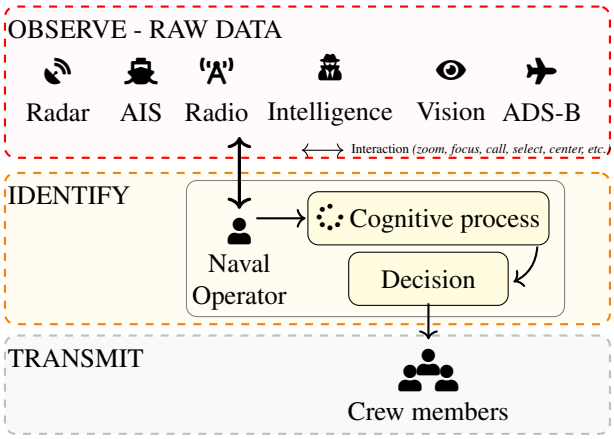

Figure 10: Current naval surveillance pipeline - the naval operator deploys cognitive processes to identify objects from all the collected information (*AIS - Automatic Identification System; ADS-B - Automatic Dependent Surveillance-Broadcast)*

A key element from field observations is that defense operators face a continuous flow of data from multiple sources; often incomplete, sometimes noisy, and frequently asynchronous [12]. In recent decades, the proliferation of sensors has led to an increase of data sources, resulting in significant cognitive burden for operators [41]. As presented in Fig 10, a naval operator may have to combine abruptly

radar detection, perform visual confirmations, monitor and engage in radio communication, coordinate with other services, and annotate observations, for each detected track in an observation-identification-transmission loop.

4.2 Ambiguity of signals

Naval operators constantly *collect* information and *correlate* fragmented signals. When data are incomplete, operators must *anticipate* developments and *complete* gaps using their expertise and knowledge. Furthermore, they are required to *prioritize* relevant tracks and signals, focusing on the most critical ones, to *make decisions* in real time, often under significant operational pressure.

In this context, information can be partial or ambiguous, which ergonomics literature refers to as situations of structural uncertainty [18]. Unlike purely procedural tasks, where each action logically follows from a clear signal [33], field observations have shown that maritime surveillance requires naval operators to interpret incomplete, unconfirmed, or contradictory signals, while remaining responsible for the potential consequences of their decisions, thus preventing the use of a rigid autonomous system.

Therefore, behind the seemingly simple task of identification, lie countless invisible actions that add up and might lead to a cognitive overload: **collect, correlate, anticipate, complete, prioritize, and make a decision**.

4.3 Human-Centric annotation

Consequently the Human-Centric annotation framework seems particularly suited to this defense context. The goal of annotation is not to replace human intervention but to enhance operators' work capacity, senses, and perception, with a purpose beyond merely recording events in a database. It must be a projection of the operator's cognitive processes, involving the selective identification of salient cues, the interpretation of signals based on task-related knowledge, and the intentional omission of elements deemed routine or unremarkable.

The Human-Centric annotation framework previously defined in Fig. 9, will be implemented alongside existing observation tasks described in Fig. 10, resulting in the dynamic collaboration presented in Fig. 11. In this setting, the naval operator can actively choose to use raw data, annotated ones, or a mixture of both to construct a decision. Consequently, the operator is active, and is not constrained by a decision from an automatic system. According to field observations, naval operators tend to reject imposed predictions from automatic systems they cannot control. Our Human-Centric annotation framework lets the operator take the final decision and stay in control. The annotation strategy is automatically constructed based on operationnal objective and context. Then each form of data can be annotated, radio communications would be classified and annotated as *acoustic* data, whereas AIS transmissions would be annotated as *numerical* data. Table 2 presents an analysis of possible annotations and demonstrates the application of the proposed annotation framework to the naval defense use case.

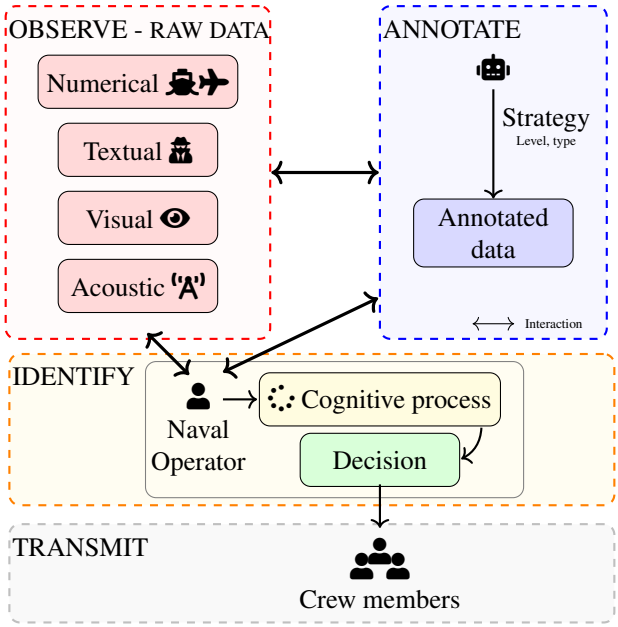

Figure 11: Proposed naval surveillance pipeline enhanced with an annotation module for each data type (textual, numerical, visual, and acoustic); the operator can use either annotations or collected information to identify objects

5 Implications and Future Directions

This section summarizes our reflections on enhancing human performance within a Human-Centric framework. An automatic system disconnected from a human operator can lead to overconfidence biases, where humans can reduce their involvement in an active analysis [31]. This phenomenon is amplified when automated interfaces leave little room for human interpretation or require scarce attention outside of critical alerts. A HITL strategy with an active engagement loop is crucial to mitigate human biases, either overconfidence or distrust, and avoid autopilot effect [17].

Additionally, gradual removal of responsibility leads to a significant decline in situational awareness, i.e., the ability to perceive, understand, and anticipate the dynamics of a changing environment [14]. This loss is precisely what needs to be avoided in the context of defense. An annotation framework where the operator is not directly providing an answer, but receives enhanced information, can be seen as a way to force human cognition [4], reducing overconfidence and distrust.

Based on observations of how maritime semaphore station and frigate operators annotate data, the *Suggestion*, *Iterative*, and *Challenger* HITL pipelines appear particularly well adapted as they preserve full human involvement.

5.1 Reframing Annotation as a Human-Centric Task

Throughout this study, the importance of human involvement in the annotation process has been repeatedly corroborated. Annotation is an intrinsically Human-Centric task, even though it is often perceived as a mechanical or aux-

Table 2: Analysis of possible annotations across modalities for naval surveillance data

Layer	**Task Name**	***[Annotation type]* - Task Description**
		Textual (Documentation - Intelligence)
Surface	Highlighting	*[Graphical]* Visually emphasize specific spans of a text.
Surface	Part-of-Speech Tagging	*[Written]* Assign part-of-speech tag (e.g., NOUN, VERB) to each token.
Surface	Lemmatization	*[Written]* Reduce inflected forms to their base form (lemmas).
Structural	Paragraph Extraction	*[Graphical]* Identify and extract text segments relevant to a query.
Semantic	Named Entity Recognition	*[Written]* Identify and label spans corresponding to entities.
Semantic	Topic Tagging	*[Written]* Assign one or more topic labels indicating thematic content.
Interpretative	Sentiment Analysis	*[Written]* Classify sentiment expressed in a sentence or document.
Interpretative	Relation Extraction	*[Written]* Identify and label semantic relationships between entities.
Interpretative	Summarization	*[Written]* Produce a condensed representation while preserving essential information.
Interpretative	Schema Generation	*[Graphical]* Transform structured or semi-structured text into an ordered representation (e.g., table, concept map).
		Visual (Camera)
Structural	Bounding Box	*[Graphical]* Delimit regions of interest in an image using rectangles.
Structural	Segmentation Mask	*[Graphical]* Assign a class label to each pixel for fine-grained object delineation.
Structural	Object Tracking	*[Graphical]* (Video) Assign identifiers to objects across frames to capture trajectories.
Semantic	Object Description	*[Graphical]* Provide structured metadata about objects (e.g., class, location, attributes).
Semantic	Classification	*[Written]* Assign labels to images or localize and label individual objects.
Semantic	Scene Detection	*[Written]* (Video) Identify shot boundaries or scene transitions.
Interpretative	Scene Graph Generation	*[Graphical]* Identify objects and their pairwise relationships (e.g., "man–riding–moto").
Interpretative	Image Captioning	*[Written]* Automatically generate descriptive captions for images.
Interpretative	Intent Captioning	*[Written]* (Video) Generate descriptions explaining goals or intentions behind actions.
		Audio (Radio)
Surface	Subtitle	*[Written]* Transcribe spoken content into aligned text segments.
Semantic	Audio Classification	*[Written]* Assign semantic categories (e.g., genre, environment, speaker identity).
Interpretative	Emotion Classification	*[Written]* Identify and label the emotional state expressed in audio.
		Numeric (AIS – ADS-B – Radar)
Surface	Highlighting	*[Graphical]* Emphasize specific numeric ranges or values (e.g., outliers).
Structural	Region of Interest Selection	*[Graphical]* Select specific segments in a time series or numerical matrix.
Structural	Data Visualisation	*[Graphical]* Project data into a space emphasizing structures (e.g., clusters, outliers).
Interpretative	Trend Classification	*[Written]* Label time series segments according to their patterns (e.g., increasing, cyclic).
Interpretative	Anomaly Classification	*[Written]* Identify and categorize abnormal data points or sequences.

iliary step. In high-stakes domains such as defense, where sensor data are abundant and real-time decisions are crucial, manual annotation alone becomes impractical. Operators often lack the time to explicitly annotate; instead, they do what may be interpreted as implicit annotation.

We claim that designing annotation systems, specifically to support human operators, can enhance decision-making capacities and reduce cognitive overload, which implies shifting the paradigm from data labeling to Human-Centered interaction.

5.2 Towards Annotation Fusion: Managing Variability and Ambiguity

A challenge that was intentionally omitted of the core scope of this paper is the question of inter-annotator variability, and more broadly, multi-agent annotation. In complex, ambiguous, or cognitively demanding situations (as in defense scenarios), annotators production may diverge significantly, even working under shared guidelines.

This subject emerged as fundamental and should be explic-

itly modeled and incorporated. For instance, annotations could be contextualized based on annotators' profiles, or different opinions could be weighted according to expertise or cognitive style. On the AI side, this also raises the issue of AI-AI variability, where multiple AI systems produce different annotations.
A use case like this one, leads naturally to the concept of annotation fusion, where diverse annotations, from humans, AI, or both, are merged to create more robust and explainable datasets. Combining multiple annotations, potentially informed by annotators' reasoning process, offers a promising direction for mitigating ambiguity, while retaining the richness of multiple viewpoints.

5.3 Annotation for decision support

Finally, multiple-criteria decision-making (MCDM) methods are well-established and powerful decision-support tools for handling numerical data. In defense applications, such systems are already used to help operators ranking and interpreting sensor outputs, as well as in selecting appropriate strategies and behaviors [3]. However, operators struggle to handle visual or textual information. Annotation could serve therefore as a complementary mechanism to enrich MCDM approaches. Conversely, MCDM techniques could also be leveraged to weight and prioritize which annotations are the most relevant.

5.4 Implementation of a Human-Centric Annotation Framework

As previously discussed, annotation in naval environments is frequently conducted under time constraints, with operational realities compelling operators to adapt beyond prescribed procedures. Furthermore, the importance of contextual information may fluctuate according to geographic and meteorological conditions. As a result, rigid systems such as rule-based algorithms or ontology-driven approaches may lack the flexibility required for such dynamic settings. A Human-Centered annotation framework should accordingly be grounded in a system that facilitates seamless interaction between human and artificial agents, supporting adaptive collaboration with the data. In this regard, multi-agent systems appear particularly promising.

6 Conclusion

Surveillance within maritime semaphore station and frigate is complex tasks involving continuous observation, identification, and decision. Currently, operators must gather information from various sources, to be correlated before making decisions. In this process annotation is implicit and lacks automatization or normalization. There are at least three possible semi-automatic annotation approaches that could improve operators' decision capacity. These approaches require to design intelligent annotation systems, beyond prescribed theoretical procedures, to meet real operational constrains and needs. Annotation resulting of human-AI collaboration, could facilitate normalization and consistency within the resulting data, while being a cognitive aid for humans operating in high-stakes environments. Data enriched semantically with meaning and context, could lead to improved machine learning models, enabling efficient identification of normal and abnormal behaviors. Future work will investigate methods for merging human insights, AI suggestions, and task-specific heuristics into a cohesive annotation pipeline.

7 Acknowledgment

This work is funded by Agence Nationale de la Recherche et de la Technologie, CIFRE n°2024/1316, within the research program of Chaire NAIADE. Authors are very grateful to maritime semaphore station and frigate operators, who answered our questions about their sea and air surveillance missions.

References

[1] Shiva Agrawal, Savankumar Bhanderi, and Gordon Elger. Semi-automatic annotation of 3d radar and camera for smart infrastructure-based perception. *IEEE Access*, 12:34325–34341, 2024.

[2] Jacob Beck, Stephanie Eckman, Rob Chew, and Frauke Kreuter. Improving labeling through social science insights: results and research agenda. In *International Conference on Human-Computer Interaction*, pages 245–261. Springer, 2022.

[3] Dragan Bojanic, Mitar Kovač, Marina Bojanic, and Vladimir Ristic. Multi-criteria decision-making in a defensive operation of the guided anti-tank missile battery: An example of the hybrid model fuzzy ahp-mabac. *Decision Making: Applications in Management and Engineering*, 1(1):51–66, 2018.

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

[5] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, pages 3–5, 2016.

[6] Daniel N Cassenti, Vladislav D Veksler, and Frank E Ritter. Editor's review and introduction: Cognition-inspired artificial intelligence, 2022.

[7] Chia-Ming Chang, Yi He, Xusheng Du, Xi Yang, and Haoran Xie. Dynamic labeling: A control system for labeling styles in image annotation tasks. In *International Conference on Human-Computer Interaction*, pages 99–118. Springer, 2024.

[8] Jing Chen, Dan Wang, Iris Xie, and Quan Lu. Image annotation tactics: transitions, strategies and efficiency. *Information Processing & Management*, 54(6):985–1001, 2018.

[9] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

[10] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. Hybrid intelligence. *Business & information systems engineering*, 61(5):637–643, 2019.

[11] Anca Dumitrache. Truth in disagreement: Crowdsourcing labeled data for natural language processing. 2019.

[12] Mica R Endsley. Situation awareness misconceptions and misunderstandings. *Journal of cognitive Engineering and Decision making*, 9(1):4–32, 2015.

[13] Mica R Endsley. From here to autonomy: lessons learned from human–automation research. *Human factors*, 59(1):5–27, 2017.

[14] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. In *Situational awareness*, pages 9–42. Routledge, 2017.

[15] Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. Llms accelerate annotation for medical information extraction. In *machine learning for health (ML4H)*, pages 82–100. PMLR, 2023.

[16] Jose Nuno Gomes-Pereira, Vincent Auger, Kolja Beisiegel, Robert Benjamin, Melanie Bergmann, David Bowden, Pal Buhl-Mortensen, Fabio C De Leo, Gisela Dionísio, Jennifer M Durden, et al. Current and future trends in marine image annotation software. *Progress in Oceanography*, 149:106–120, 2016.

[17] Jean-Michel Hoc. Towards ecological validity of research in cognitive ergonomics. *Theoretical issues in ergonomics science*, 2(3):278–288, 2001.

[18] Erik Hollnagel. *The ETTO principle: efficiency-thoroughness trade-off: why things that go right sometimes go wrong*. CRC press, 2017.

[19] Remi Kalir and Antero Garcia. Annotation guidelines. *Annotation*, 2019.

[20] Yusuke Kamoi, Yosuke Furukawa, Tatsuya Sato, Yuya Kiwada, and Tomohiro Takagi. Automatic image annotation based on visual cognitive theory. In *NAFIPS 2007-2007 Annual Meeting of the North American Fuzzy Information Processing Society*, pages 239–244. IEEE, 2007.

[21] Hannes Kath, Thiago S Gouvêa, and Daniel Sonntag. A human-in-the-loop tool for annotating passive acoustic monitoring datasets. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 341–345. Springer, 2024.

[22] Gary Klein, Brian Moon, and Robert R Hoffman. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems*, 21(5):88–92, 2006.

[23] Jan-Christoph Klie, Richard Eckart De Castilho, and Iryna Gurevych. From zero to hero: Human-in-the-loop entity linking in low resource domains. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6982–6993, 2020.

[24] Iuliia Kotseruba and John K Tsotsos. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 53(1):17–94, 2020.

[25] Sushant Kumar, Sumit Datta, Vishakha Singh, Sanjay Kumar Singh, and Ritesh Sharma. Opportunities and challenges in data-centric ai. *IEEE Access*, 12:33173–33189, 2024.

[26] Sang-Heon Lee, Hae-Gwang Park, Ki-Hoon Kwon, Byeong-Hak Kim, Min Young Kim, and Seung-Hyun Jeong. Accurate ship detection using electro-optical image-based satellite on enhanced feature and land awareness. *Sensors*, 22(23):9491, 2022.

[27] Shixia Liu, Changjian Chen, Yafeng Lu, Fangxin Ouyang, and Bin Wang. An interactive method to improve crowdsourced annotations. *IEEE transactions on visualization and computer graphics*, 25(1):235–245, 2018.

[28] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6122–6131, 2019.

[29] Ayodele Marvellous, Bamidele Matthew, Mauro Pezzè, Silvia Abrahão, and Birgit Penzenstadler. Human-in-the-loop ai engineering: Enhancing collaboration between developers and end users. 2025.

[30] Nick Pangakis and Sam Wolken. Keeping humans in the loop: human-centered automated annotation with generative ai. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1471–1492, 2025.

[31] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[32] Jan L Plass, Roxana Moreno, and Roland Brünken. *Cognitive load theory*. Cambridge university press, 2010.

[33] Jens Rasmussen. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE transactions on systems, man, and cybernetics*, (3):257–266, 2012.

[34] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2121–2131, 2015.

[35] Mary B Ruvane. Defining annotations: a visual (re)-interpretation. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–5, 2006.

[36] Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, et al. Is one annotation enough?-a data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215–33232, 2022.

[37] Anastasios Stamoulakatos, Javier Cardona, Chris McCaig, David Murray, Hein Filius, Robert Atkinson, Xavier Bellekens, Craig Michie, Ivan Andonovic, Pavlos Lazaridis, et al. Automatic annotation of subsea pipelines using deep learning. *Sensors*, 20(3):674, 2020.

[38] Teodor Stoev, Kristina Yordanova, and Emma L Tonkin. Experiencing annotation: Emotion, motivation and bias in annotation tasks. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 534–539. IEEE, 2023.

[39] Jinhui Tang, Qiang Chen, Meng Wang, Shuicheng Yan, Tat-Seng Chua, and Ramesh Jain. Towards optimizing human labeling for interactive image tagging. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 9(4):1–18, 2013.

[40] Dejan Todorovic. Gestalt principles. *Scholarpedia*, 3(12):5345, 2008.

[41] Christopher D Wickens. Multiple resources and mental workload. *Human factors*, 50(3):449–455, 2008.

[42] Christopher D Wickens, Sallie E Gordon, Yili Liu, and J Lee. *An introduction to human factors engineering*, volume 2. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[43] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.

[44] Dengsheng Zhang, Md Monirul Islam, and Guojun Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362, 2012.

[45] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020.