A User-Centered Ontology for Explainable Artificial Intelligence in Aviation

Denys Bernard[1], Jaume R. Perello-March[2], Ignacio Solis-Marcos[2] and Martine Cooper[2]
[1]*Airbus SAS, Architecture and Integration, 1IVA, France*
[2]*Airbus SAS, Human Factors & Ergonomics in Design, 1YDN, France*

Keywords: Artificial Intelligence, Aviation, Explainability, Ontology.

Abstract: Aviation authorities will require future decision assistance systems based on Artificial Intelligence (AI) to be explainable in order to enhance trust, safety, situation awareness (SA) and to promote appropriate use of the system. We anticipate that authoring implementable explainability requirements will be a challenge for the relevant stakeholders in the aerospace industry. Here we propose an ontology for explainable AI (XAI) from a user centered perspective in aviation. We propose that the development of an adequate mental model of XAI has to be considered as a naturally dialogic process between the user and the AI, where the need for an explanation can be approached as a question. The explanation specification process describes the informational content of explanations,, whose main components are an *Explanans* and an *Explanandum* linked by the appropriate *discourse relation*. The *Explanandum* denotes the aspect of the outcome of the system about which the operator needs an explanation, and the *Explanans* is typically a set of true facts which actually satisfy the operator's cognitive need. We understand explanation as a communication act with the purpose of making the user accept or better understand the *Explanandum*. Thus, an explanation is successful only if a particular relationship holds between the explanans and the explanandum. To understand such a relationship, we refer to the theory of discourse where the so-called *discourse relations* act as the logical core of a discourse, and are constitutive of its consistency. By using this ontology, aviation Human Factors and operations practitioners will be able to specify the content of explanations in order to maximize the acceptability and usability of explanations by the user.

1 INTRODUCTION

Artificial Intelligence (AI) is bringing a vast set of new potential applications and solutions for the aviation industry including aircraft design, operations, production, maintenance, environment, and air traffic management, to name a few (EASA, 2023a). Human Factors for AI are among the certification requirements that aviation authorities will demand to certify future AI-based systems, (EASA, 2023b). These include:

- AI operational explainability as "the capability to provide the human end users with understandable, reliable and relevant information with the appropriate level of details and with appropriate timing on how an AI/ML application produces its results".
- And human-AI teaming "to ensure adequate cooperation or collaboration between human end users and AI-based systems to achieve certain goals".

Such interest in ensuring the development of explainable AI (XAI) is not trivial. Extensive work has been exploring this notion in the recent past years since AI models can lead to unpredictable outputs that may be difficult to explain by the end users, and thus, hamper the human-AI teaming (Druce et.al., 2019; Endsley, 2023). This can lead to operators making wrong mental models about the AI, which are then difficult to break. Hence, explainability is fundamental for developing an accurate mental model (Druce, 2019).

According to Ensdley (2023), "explainability pro-vides information in a retrospective manner, describing the logic, pro-cess, factors, or reasoning on which the system's actions or recommendations were based". In a nutshell, XAI refers to WHY the system did something, in terms of its capabilities and

Bernard, D., Perello-March, J., Solis-Marcos, I. and Cooper, M.
A User-Centered Ontology for Explainable Artificial Intelligence in Aviation.
DOI: 10.5220/0012978600004562
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 2nd International Conference on Cognitive Aircraft Systems (ICCAS 2024)*, pages 88-94
ISBN: 978-989-758-724-5

processes. Another often related concept is transparency, which refers to WHAT is the system doing now and in the near future. Both contribute to creating adequate trust in the system and support situation awareness by ensuring the predictability of the system behavior (Endsley, 2023).

The relevance of an explanation depends on cognitive, social and operational aspects of the current situation. Specifying what would be a good explanation is a multi-disciplinary task, which should result in implementable technical requirements. There is a need for a set of concepts to communicate efficiently on explainability, in particular, to author understandable and verifiable technical requirements.

We intend to clarify this by formalizing the system explainability concepts into an ontology. According to Keller (2016), ontological techniques were initially developed in Artificial Intelligence to handle the knowledge used and processed by intelligent agents in performing reasoning tasks. But those methods have spread over a variety of domains, including the Semantic Web and the design of data exchange formats. Languages to represent ontologies have been standardized and widely adopted, in particular OWL (Ontology Web Language). Formally, an ontology is a set of statements that describes classes of concepts by their interdependencies, in particular the relationships that must hold between the class instances.

The first problem to be addressed when clarifying the concepts about explainability, is that "explanation" is polysemic in our daily language. Consider the following examples from the Collins dictionary (Collins, 2024). "Explanation" in each of them refers to a different high level concept:

- Explanation as a dialog: "Forget about explanations; they'd only end in arguments".
- Explanation as a speech act: "It is of no use to attempt an explanation".
- Explanation as a logical construct: "There is a simple cognitive explanation as to why numbers get blurry after three".
- Explanation as a text: "There is a lengthy explanation about the pros and cons."

In Walton (2004), explanation is analyzed as a speech act -possibly a complex one- whose main objective is to "transfer understanding" as an answer to a question. Although we reserve "explanation" to refer to the logical content of such speech acts, our approach is compliant with the major ontological choices made in Walton (2004): explanation acts are understood as parts of explanation dialogs in which questions specify the needed explanatory content. To assemble a coherent toolbox of operational concepts, we articulated concepts from diverse sources into a single dedicated ontology. We considered not only specialized ontologies of explanation such as Walton (2004), Chari et al., (2020) and Lindner (2020), but also fundamental ontologies (Borgo, 2022), and publications about speech acts and discourse theories (Green, 2021; Smith, 2015). We first remind the notion of communication acts (section 2), which we need to clarify the variants of "explanation"; then we detail the logical structure of explanations (section 3); next we introduce the dialectical structure of explanations (section 4), which has certain practical outcomes; finally we conclude with methodological considerations (section 5).

2 COMMUNICATION ACTS

Whereas EASA (2023a) refers to "explainability" as a capability, "explanation" is defined as "information [...] on how an AI application produces its results". However, "information" is in itself polysemic (Smith, 2015): is it a communication act? a piece of knowledge? an information bearer? (i.e., a sequence of symbols representing logical content, such as text, utterances, images). In this section, we suggest to keep "explanation" to refer to the logical content to be provided to the operator. Nevertheless, the concept of explanation as a speech act will also play a central role.

To articulate the social, logical and cognitive dimensions of explanation, we refer to *Communication Acts*, derived from the speech act theory (Green, 2021; Smith, 1984), as it has been adapted to communication agents in information systems (Boella, 2007; Ferrario, 2007). Figure 1 shows a simplified view of the concepts and relations about speech acts.

Concept names start with uppercase letters (e.g. *Agent*). Concepts and relations prefixed by "*dolce*:" are borrowed from the fundamental ontology DOLCE (Borgo, 2022). Roles like *sender* or *receiver* are depicted here as relations for simplicity, but they should be modeled following more rigorous representation rules. On the representation of roles see for example Vieu et al. (2008).

Communicative cognitive agents are supposed to be animated by *Mental attitudes.* Boella (2007) associates *Mental attitudes* to roles. *Mental attitudes* of the roles involved in communication acts are manifested by the performance of the communication act. Two sub classes of *Mental attitudes* are needed for the formalization of communication acts: *Goals*

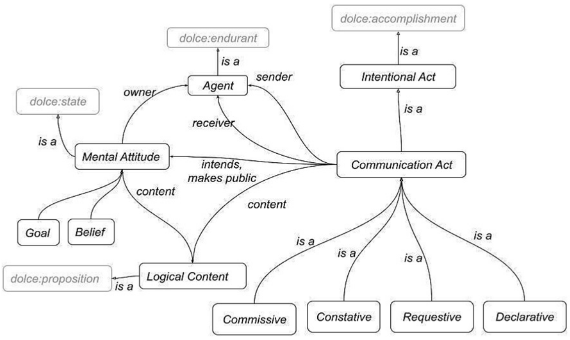

Figure 1: Communication acts.

and *Beliefs* (Boella 2007). *Mental attitudes* are defined by their type (*Goal* or *Belief*) and *Logical.*

Content. The *Logical content* is made of truth-evaluable propositions (it inherits from the DOLCE *proposition* concept). Different *Mental attitudes* may have similar content, for example: I believe the door is closed, or I intend to close the door. Both intentional states have the same content ("the door is closed") but differ from their nature (belief and goal).

The purpose of *Communication Acts* is to make the *receiver* adopt a specific *Mental attitude*. A speech act is determined by its "force" and its "content" (Green, 2021). We will assume that the logical *content* of the *Communication Act* is similar to the propositional *content* of the intended *Mental attitude* (Green, 2021). This last statement is an oversimplification in the perspective of pragmatic theories of communication (Recanati, 1998): the intended cognitive effect of a speech act can overcome its strict logical content by triggering intentionally further inferences (or "implicatures" (Wayne, 2024)). A *Communication Act* can be decomposed into several subclasses associated to different "features", correlated to the nature of the intended *Intentional State* of the addressee: *Constative Acts* aim at updating beliefs, while *RequestiveActs* intend to change the goals of the addressee, while a *Commissive act* is about the intentions of the sender himself. By saying "the door is closed", the speaker intends to make the addressee believe that the door is closed. Whereas by saying "please, close the door" the speaker aims at making the addressee to close the door In particular, *Constative Acts* are emitted to make the receiver believe their content. *Explanation acts* will be understood in the next section as *Constative acts*, which follow (or explain) a primary act, which can be a *Constative act* (when the system delivers information), or a *Directive act* when the system is a recommender.

3 THE CONTENT OF EXPLANATIONS

We define *Explanation Acts* as a specialization of *Constative* communication acts. We have already mentioned that the word "explanation" may refer to different object categories in usual language. For the sake of clarity, we propose an ontological choice, that the concept *Explanation* is dedicated to the logical content of *Explanation Act*s, as depicted in Figure 2:

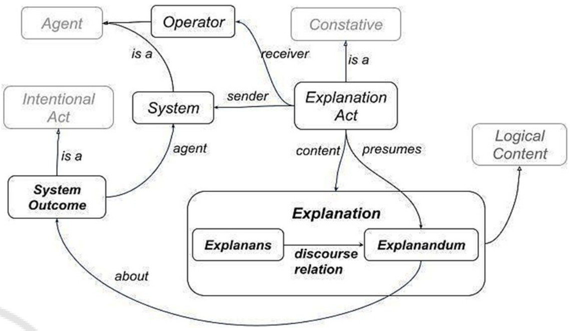

Figure 2: The content of explanations.

The scope of this ontology is on our current use case: systems' AI explainability. Hence, the *sender* of an Explanation Act is the system of interest (i.e., the machine), and the *receiver* of an *Explanation Act* is the human *Operator* (i.e., the pilot). The structure of an *Explanation* is a construct of three elements: the *Explanans*, the *Explanandum,* and their link, a *discourse relation* (Hovy & Maier, 1992; Mann & Thompson, 1988). The *Explanation,* the *Explanans* and the *Explanandum* are *Logical Contents*, i.e., they are all truth- evaluable elements (they are inherited from DOLCE's *proposition*). For example, the statement "The airplane turns right to avoid a cloud" links the *Explanandum* "The airplane turns right", to the *Explanans* "to avoid a cloud", with a type of discourse relation named *VOLITIONAL RESULT* according to the taxonomy from Hovy & Maier (1992). This kind of relation links a desired effect to the event or action which could cause this effect (we further disclose the taxonomy of discourse relations that we intend to use for system explainability in Figure 4).

The *Explanandum* is about aspects of the *System Outcome*, hence, it is generally known, or presumed by both agents (Walton, 2004). The *Explanans* and the *Explanandum* must be related by the appropriate discourse relation for fulfilling the cognitive need of the operator. Different types of discourse relation between the Explanans and the Explanandum define different explanation subclasses. A complete taxonomy of discourse relations is given in Hovy & Maier (1992), and definitions of the most useful discourse relations for explanation purposes is available in Mann & Thompson (1988). To summarize, for an *Explanation Act* to be successful, it must include the following necessary conditions:

- The *Explanandum,* which is an aspect of the *System Outcome*;
- The *Explanans*, which describe actual facts, already known or not by the operator, but which can be accepted.
- The *discourse relation*, which links the *Explanans* to the *Explanandum*.

For example, in causal explanations, *Explanans* describe the causes of the *Explanandum*: "The weather is degrading because atmospheric pressure is decreasing" (i.e., *NonVolCause* in the taxonomy from Hovy & Maier, 1992).

The success of an *Explanation* depends on whether the operator acknowledges or agrees with the proposal. Particularly, it requires that the *discourse relation* in the *Explanation* is of an appropriate type. For example, "the airplane turns right, it has been designed to make this kind of movement" is acceptable from a strict logical viewpoint, but is probably not an acceptable explanation for the pilot in an operational situation. That is, because the relationship between the two segments of the sentence is not explanatory. More generally, *Discourse relations* are central in theories of discourse because the consistency of the discourse depends on them: they are sometimes presented as "the glue" of discourse (Asher & Lascarides, 2003). As the *Explanandum* is presumed by the Explanation Act -"presumption condition" in Walton (2004)-, the main novelty in an *Explanation Act* is often the relation itself between the *Explanans* and the *Explanandum*. The *Explanans* itself may be known or not before the *Explanation Act*.

The type of *Explanation*, and in particular its discourse relation, must be tailored to the operator's cognitive expectations. Research suggests that humans have a preference for contrastive explanations (Miller, 2019). It means that operators expect not only an explanation of why something happened, but why something happened rather than something else. For example, after a recommendation to turn right, instead of asking (Q1) "*why should I turn right?*" the pilot might ask (Q2) "*why should I turn right rather than turning left?*". The explanations required by (Q1) and (Q2) have different *Explanandum* types, then the discourse relations will also be different: the *Explanans* for (Q1) could be a desired effect of turning right, whereas (Q2) appeals for a justification of the preference order. In addition, explanations must be contextual and adapted to the specificities of each particular scenario (Druce et.al., 2019).

4 EXPLANATION AND DIALOG

In daily conversations, an explanation is also a type of dialog. Walton (2004) has explored this idea through a theory of explanation based on an archetypal dialog, which starts by a question. The dialog succeeds when understanding is transferred to the questioner. The cognitive need is specified by the opening question of the explanatory dialog. In the case of systems' explainability, the interaction does not necessarily take the form of a dialog, but the way of describing the need for an explanation by a question remains efficient. The need for an explanation is understood by determining what question the user could have asked. This method is not too restrictive because we focus on the logical content of explanations, and, as noted by Walton (2004), explainability is inherently dialogic. The most simple, although not unique, explanatory dialog pattern takes the form of a question-answer dialog, supported by the primary outcome of the system of interest, or the description (declaration) of what the system intends to do:

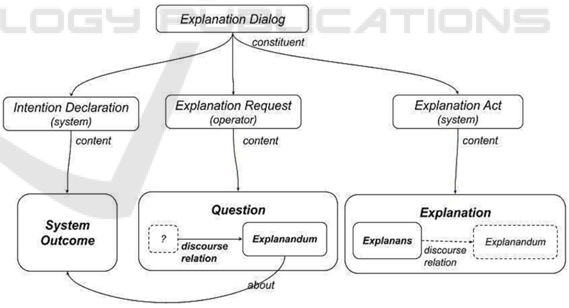

Figure 3: The basic explanation dialog.

The content of the *Explanation* is built progressively through the dialog. The part of the *Explanandum* to be explained and the acceptable discourse relations that would make an acceptable explanation are set by the question. In particular, as noted in Miller (2019), explanations are often contrastive: the operator needs to understand why the outcome of the system is what it is rather than something else. Take the example of a diversion assistant which recommends to re-route to either *Bordeaux, Toulouse, or Agen.* By asking: "*Why is Bordeaux preferred to Toulouse?*"; the pilot requires an explanation, whose *Explanandum* is: *"Bordeaux is preferred to Toulouse*" (which is part of the system outcome), and the discourse relation is a *JUSTIFICATION*. Following we propose an example of how this dialog would occur:

-SYSTEM: "*Re-routing options are Bordeaux, Toulouse, or Agen"*.

-PILOT: "*Why is Bordeaux preferred to Toulouse?*"

-SYSTEM: *"Bordeaux is preferred to Toulouse in order to save fuel"*

Finally, the *Explanans* given in the last utterance justifies that diverting to Bordeaux is more operationally convenient. Based on the taxonomy of *discourse relations* established in Hovy & Maier (1992), Figure 4 shows the most frequent discourse relations which appear in explanations, as well as some examples of questions that the operator could have asked to obtain an acceptable explanation.

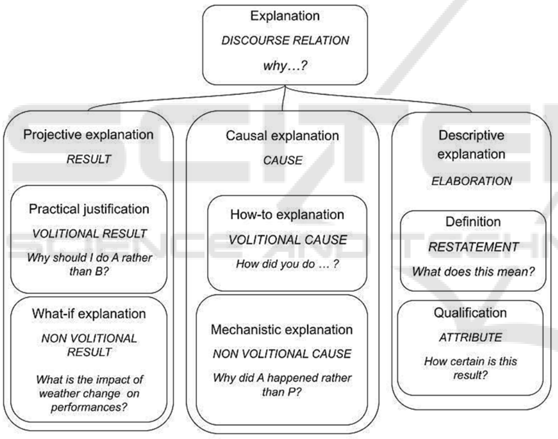

Figure 4: A simple taxonomy of explanation.

The taxonomy of discourse relations proposed in Hovy & Maier (1992) makes a complete inventory of discourse relations used in the literature. It distinguishes discourse relation which expresses a conceptual relation between the two segments (*IDEATIONAL* relations) and the relation which expresses directly a communicative intent of the speaker (*INTERPERSONAL* relations). *INTERPERSONAL* relations link two sentences in order to obtain a cognitive effect. The example ontology on Figure 4 uses only *IDEATIONAL* relations, because one of our methodological goals is to make explainability requirements as objective and verifiable as possible.

For example, the *IDEATIONAL* category includes relations for causality and part-whole relations. *INTERPERSONAL* relations include for example *JUSTIFICATION* or *MOTIVATION.* In practice, the relation between two segments in an explanation can belong to both categories. For example, in "You should climb to flight level 300, that will save fuel", the two segments of the sentence are linked causally, but at the same time, the purpose of the statement is to convince the addressee that climbing would be a good decision. When authoring requirements, we should prefer explanations defined in *IDEATIONAL* terms rather than *INTERPERSONAL* ones. This principle would reinforce the implementability and the verifiability of requirements. The duty of the Human Factor practitioners would be to translate interpersonal goals (convince, transfer understanding, build trust), into the terms of information semantics, including what Hovy & Maier (1992) classified as *IDEATIONAL* relations.

5 A USER CENTERED XAI ONTOLOGY

Our ontology proposes a method for defining the content, the dialog structure and a taxonomy for different types of explanations for AI in aviation.

Previous related work from Sutthithatip et al. (2021) has reviewed different ways to implement XAI in aviation to support designers, pilots, air traffic controllers and maintenance operators at several levels by:

(1) extracting and integrating the information,
(2) understanding the situation,
(3) predicting the outcomes or consequences of actions to make a decision, and
(4) implementing the desired action course

Essentially, this model proposes that XAI can support aviation stakeholders in achieving a good situational awareness (Levels 1-3) and then in choosing and implementing an adequate course of action. If done appropriately, XAI can contribute to mitigate mental overload, enhance human-machine teaming performance and overall operational safety.

We expect our proposal to help practitioners identify and define the appropriate explanations for each particular use case or scenario.

Providing support in the decision making process is one of the key areas where XAI can make a substantial contribution to make aviation operations safer. Decision-makers in aviation are often faced with time-constrained and safety-critical decisions, for which, having accurate information at the right time is essential. However, it is well-known that humans suffer from several decision biases (Kahneman & Klein, 2009) that may lead to wrong decisions, particularly in emergency situations, or when the decision-makers lack experience, expertise or time to come up with an analytical decision.

XAI has a huge potential in supporting aviation operators by mitigating decision errors. However, to achieve that, it has to provide the information in a self-explanatory way so operators can comprehend the situation and the consequences or their decisions. Therefore, we expect our methodology will support describing how to make and what is the necessary content of a good explanation.

6 CONCLUSIONS

We proposed an ontology summarizing the main elements for XAI in aviation. In this ontology, we acknowledge the dialectical dimension of explanations (Walton, 2004), and we frame it upon speech acts theories. We also acknowledge that discourse theories are relevant for understanding the rhetorical structure of explanations. In particular, an explanation is understood as a logical structure with three terms: (1) an *Explanandum*, which is an aspect of the outcomes of the system of interest, (2) the *Explanans*, and (3) the discourse relation which links the *Explanans* and the *Explanandum*.

Our main contribution is to enhance the role of discourse relations to make explanations successful. The concepts defined here will serve as foundations for explainability requirements in the early phases of systems development. Good quality explainability requirements should translate cognitive needs or concerns into implementable and verifiable design principles. Those requirements will be the point of contact between practitioners in charge of capturing the cognitive needs of the operators, and engineers in charge of designing the system. Our assumption is that explainability requirements will be implementable and verifiable if they rest on the logical structure of information, as managed by the system whose outcomes or recommendations are to be explained. On the theoretical side, a lot of work remains to be done. In particular, building the map between a taxonomy of cognitive needs to be fulfilled through explainability and the corresponding types of explanation. For this, a further formalization effort might be needed regarding the taxonomy of discourse relations and their semantics.

REFERENCES

Asher, N., Lascarides, A., 2003, Logics of Conversation. Cambridge University Press.

Boella, G., Damiano, R., Hulstijn, J., van der Torre, L., 2007, A common ontology of agent communication languages: Modeling mental attitudes and social commitments using roles. Applied Ontology. 2. 217-265.

Borgo, S. , Ferrario, R., Gangemi, A., Guarino, N., Masolo, C., Porello, D., Sanfilippo, E. M., & Vieu, L., 2022, DOLCE: A descriptive ontology for linguistic and cognitive engineering. Applied ontology, 17(1), 45-69.

Chari, S., Seneviratne, O., Gruen, D.M., Foreman, M.A., Das, A.K., McGuinness, D.L., 2020. Explanation Ontology: A Model of Explanations for User-Centered AI. In: J. Z. Pan et al. The Semantic Web – ISWC 2020. ISWC 2020. Lecture Notes in Computer Science, vol 12507. Springer

Collins, 2024. Collins Dictionary entry on "explanation", consulted on line in April 2024.

Druce, J., Niehaus, J., Moody, V., Jensen, D., & Littman, M. L., 2021. Brittle AI, causal confusion, and bad mental models: challenges and successes in the XAI program. arXiv preprint arXiv:2106.05506.

EASA, 2023a. Concept Paper, First usable guidance for Level 1&2 machine learning applications, European Union Aviation Safety Agency, issue 2, Feb 2023.

EASA, 2023b. ARTIFICIAL INTELLIGENCE ROADMAP 2.0. Human-centric approach to AI in aviation, European Union Aviation Safety Agency, May 2023.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human factors, 37(1), 32-64.

Endsley, M. R., 2023. Supporting Human-AI Teams: Transparency, explainability, and situation awareness. Computers in Human Behavior, 140, 107574.

Ferrario, R., Prévot, L., 2007. Formal ontologies for communicating agents. Applied Ontology. 2. 209-216.

Green, M., 2021. Speech Acts. In E. N. Zalta (ed.), The Stanford Encyclopedia of Philosophy

Hovy, E.H., Maier, E., 1992. Parsimonious or profligate: how many and which discourse structure relations? University of Southern California. Information Sciences Institute, ISI Research report.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. American psychologist, 64(6), 515.

Keller, R. M. (2016) Ontologies for aviation data management, 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 2016, pp. 1-9

Lindner, F., 2020. Towards a Formalization of Explanations for Robots' Actions and Beliefs. Proceedings of "RobOntics: International Workshop on Ontologies for Autonomous Robotics", JOWO 2020

Mann, W., Thompson, S., 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. Text. 8. 243-281.

Miller, T, 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Recanati, F., 1998. Pragmatics. Routledge Encyclopedia of Philosophy. 7.

Smith, B.. 1984. Ten Conditions on a Theory of Speech Acts. Theoretical Linguistics. 11. 311-330.

Smith, B., Ceusters, W., 2015. Aboutness: Towards Foundations for the Information Artifact Ontology. Proceedings of the International Conference on Biomedical Ontology, ICBO 2015, Lisbon

Suthathip, S., Perinpanayagam, S., Aslam, S., & Wileman, A. (2021, October). Explainable AI in aerospace for enhanced system performance. In 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC) (pp. 1-7). IEEE.

Vieu, L. & Borgo, S., & Masolo, C., 2008. Artifacts and Roles: Modeling Strategies in a Multiplicative Ontology. Frontiers in Artificial Intelligence and Applications, 183 (1), 121-134.

Walton, D., 2004. A New Dialectical Theory of Explanation. Philosophical Explorations. 7(1)

Wayne, D., 2024. "*Implicature*", The Stanford Encyclopedia of Philosophy (Spring 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.)