

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
L.H.	8	Contents	Part 6	ed	When there is only one subsection it feels kind of useless especially when the section title is already "Overview"	Delete 5.1 General	
L.H. Ph.D.	12	Contents	Part 6	ed	In all subsections, explainability is between quote marks except here. Why? I also believe that the choice of versus (vs) implies that explainability is different from (or opposed to) the other term while there is some overlapping.	Change for something like: 6.1 General 6.2 Explainability 6.3 Explainability and Interpretability 6.4 Explainability and Transparency 6.5 Explainability and AI Literacy 6.6 Explainability and Evaluation 6.7 Explainability and development tools 6.8 Explainability and task specifications	
Ph.D.	136	Introduction	§2	te	Explainability is not the only one method to evaluate trustworthiness.	Change for: While the overarching goal of explainability is to contribute to evaluate the trustworthiness of AI systems	
L.H.	138-145	Introduction	§2	te	It misses an example of one of the main goals of explainability: failure analysis	Add: For regulators, it enables failure cases analysis of embedded AI systems.	
L.H.	146-148	Introduction	§3	te	It is a very light description of the work done here	Add: This document proposes laying down a taxonomy of "Explainability" in the context of AI (XAI). Besides, it will specify related concepts and ideas often mixed up with XAI to prevent any confusion.	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
Ph.D.	158	1		ge	Isn't it important to mention healthcare laboratories and professionals?	Add: ...not limited to ... healthcare laboratories and professionals, ... and users.	
L.H.	181	3.2		ed	Why a capital letter here when all others do not have one?	Change for: interpretability	
L.H.	182-183	3.2		te	This definition does not mention that it is the algorithm communicating. It should also specify that depending on the audience objectives an algorithm can be interpretable while it is not for another with a different objective (ex: an audience of developers for debugging vs an audience of regulators)	Suggest to change it to: <algorithms> ability to communicate the reasons for an AI system's behaviour convincingly to a targeted human audience. Add: Note 2: It should be noted that an algorithm can be interpretable for a specific audience (ex: ML practitioners) while remaining opaque for another (ex: regulators, end-users ...) Note 3: Reader should note that being convincing does not necessarily mean being "right".	
Ph.D.	182-183	3.2		te	I think it is the definition of understandability not interpretability. Interpretability is a comprehension that have to make sense for the human. It is the capability to exploit this comprehension in its own application domain. In other words, interpretability is the capability to project the user understanding in its own application domain. Interpretation is a mental process that allows to	Change for: Understandability Add: Interpretability: relates to the capability of an element representation (an object, a relation, a property...) to be associated with the mental model of a	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
					<p>pass from a technical understanding of an element to its use in a specific application domain.</p> <p><i>Cf. R. S. Michalski, A theory and methodology of inductive learning, in Machine learning, Springer, 1983, pp. 83–134</i></p> <p><i>Cf. [DEEL project] Mamalet, F., Jenn, E., Flandrin, G., Delseny, H., Gabreau, C., et al. (2021). White Paper Machine Learning in Certified Systems. Research report, IRT Saint Exupéry ; ANITI.</i></p>	<p>human being.</p> <p>[SOURCE:DEEL Project]</p>	
L.H. Ph.D.	188-190	3.3		te	For most AI models, it is not the <i>AI system</i> that expresses important factors but external methods that reveal the important factors influencing the <i>AI system</i> . In addition, the part “in a way that human can understand” belong to interpretability.	<p>Suggest to change it to:</p> <p>Ability to reveal the important factors influencing the AI system results.</p>	
L.H.	207-208	3.5		te	It should be mentioned that the explanatory information should be relevant for the stakeholder needing it. Explanations should not be the same depending on your job and/or functions	<p>Suggest to change it to:</p> <p><policy> ability to provide stakeholders of an AI system with relevant explanatory information beyond the AI system's results, that is meaningful for the stakeholders.</p>	
L.H.	253	3.10		ed	“global” by itself does not make any sense.	<p>Change it to:</p> <p>global explanation</p>	
Ph.D.	254	3.10		te	“global/local” interpretation does not make sense.	<p>Remove:</p> <p>or an interpretation</p>	
L.H.	260	3.11		ed	“local” by itself does not make any sense.	<p>Change it to:</p> <p>local explanation</p>	
Ph.D.	254	3.10		te	“global/local” interpretation does not make sense.	<p>Remove:</p>	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
						or an interpretation	
L.H.	269	3.12		ge	“means” is a little been vague	Replace means with methods	
L.H. Ph.D.	275	3.13		te	It would be nice to emphasize that an input feature for a model is not necessarily the inputs a user gives as entry as several layers of processing could be applied	Add: Note 2 to entry: an input feature for a model is not necessarily the inputs a user gives as entry as several layers of processing could be applied	
L.H.	277	3.13		ed	Closing quotes but no opening ones	Remove it	
L.H.	307-308	3.18		te	Unclear even for someone working in the field	Suggest to change it to: <explainability> property of a method that aims at providing explanations concerning an <i>AI system</i> without any priors concerning this system.	
A.P.	307-308	3.18		te	I do not agree with Lucas' comment. For me, a third definition on black-box could be added regarding a black-box method.	Add: Black-box <explainability method> method that can be applied to opaque box. In opposition to white-box. White-box <explainability method> method that requires access to the model weights, gradients, and/or architecture.	
L.H. Ph.D.	311	3.19		te	I would avoid the use of interpretability. In my opinion, interpretability relates to the stakeholders and changes depending on their end goals. However, when one is talking about intrinsically explainable models it means that the design of the model holds explanations but it does not assume its intelligibility.	Replacing it with one of these terms: intrinsic explainability explainable by-design (<i>preferred</i>) intrinsically explainable	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
					Cf. remarks about explainability and interpretability (§3.2, 3.3)	And change the definition to: property of an AI model that holds its criteria and decision process in its structure or content	
L.H.	318	4		ge	It misses abbreviations already uses previously but that should be mentioned again.	Replace XAI with eXplainable Artificial Intelligence Add AI Artificial Intelligence ML Machine Learning	
L.H. Ph.D	323	5.1		ed	Unnecessary subsection numbering	Remove it	
L.H.	324	5.1		ed	Emphasize the “why”	Change for: “The basic goal [...] understand WHY an AI...”	
Ph.D.	324	5.1		te	It would be better to talk about “decisions” than “predictions” to be generic.	Replace “precisions” with “decisions” in the entire paragraph.	
Ph.D.	324	5.1		ge	This overview is about XAI which I believe is a limitation of the scope of this document including comprehensibility and intelligibility as objectives.		
L.H. Ph.D.	322-331	5		ge	Before reading the section, the topic of the overview is unclear, too vague. Renaming this part Overview of XAI would be beneficial for clarity. However, that would imply moving this section closer to section 11.	Suggest moving this section to 11.1 and replacing it with: The basic goal of XAI is to enable stakeholders to understand why an AI system produced its decisions. XAI has been emerging to attempt deciphering the internal workings of <i>AI systems</i> in a human-readable way. Objectives of this field include (but are	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
						<p>not limited):</p> <ul style="list-style-type: none"> - Explaining an incorrect decision of a model for a given input. - Ensuring, that a decision was taken for the <i>right reasons</i>. - Detecting bias in the training data that might be, or not, be harmful to the decision process. - Strengthening the confidence one can have in the system. <p>The relevance of those objectives depends on the stakeholders and what they are trying to achieve. They can be interested in achieving one or several of them.</p>	
L.H. Ph.D.	350-352	6.1	§3	ed	<p>From “Second, it includes...” to “... of a given factor”. This does not mean anything or at least it is not understandable.</p> <p>What are the definitions of causes and factors?</p>	Cannot suggest something here as the message is unclear	
L.H.	367-369	6.2	§1	ed	<p>It is strange to say that this document will not “enforc[e] the approach used to achieve that effect” while in the scope (Clause 1) it is stated that: “This document describes approaches and methods that can be used to achieve explainability objectives of stakeholders with regards to ML models and AI systems’ behaviours, outputs, and results”.</p>	Clarify either the Scope or this claim	
L.H. Ph.D.	381	6.2	First list, first item	te	The term ‘content’ is strange	Replace content with items	
L.H.	381-417	6.2	All listed elements	ed	Those definitions should be included in section 3		

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
Ph.D.	398, 400	6.2	First list, two last items	ed	Here are the definitions needed to understand 6.1§3		
Ph.D	417	6.2	Second list, last item	te	<p>I am not completely agree. As I said about interpretability, an interpretation is the projection of the understanding of the model by the human in its own application domain.</p> <p>This last item as written can be a definition of comprehension. However, the comprehension of an explanation need further elements as outputs, inputs and others depending the context.</p>	<p>Change for:</p> <ul style="list-style-type: none"> - The comprehension is the result of understanding (by a human) a given explanation. It is a mental result specific to each human in human audience. <p>Add:</p> <ul style="list-style-type: none"> - An interpretation is the projection of the understanding of the model by a human in its own application domain. It is a mental process specific to each human. 	
Ph.D	418	6.2		te	As interpretation is a human process, explainability methods aims explanations and only explanations.	<p>Change for:</p> <p>“Explainability” methods aim at producing explanations.</p>	
Ph.D.	421-425	6.3	note	ge	<p>My point of view in summary:</p> <p>An explanation is produced by algorithms that analyze the model decision. It highlight one or more elements (factor or cause) of the AI system behaviour in the decision-making process; never the whole. These algorithms are explainability methods post-hoc or model by design.</p> <p>Comprehension is a human mental process that allows to a person to analyze an explanation of the decision making eventually according to further contextual elements (inputs, outputs, etc.). This analysis allows it to understand the AI system decision. This understanding is mentally expressed in an AI system linguistic corpus.</p> <p>Interpretation is the projection of the understanding of the AI system decision -and its decision making- by a human in its own</p>		

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
					application domain. This interpretation is mentally expressed in the application domain linguistic corpus.		
L.H.	448-451	6.4	§4	ge	The two “Being able to...” do not bring any additional information for this specific part and they are redundant.	Remove them	
Ph.D.	427-457	6.4	all	ge	Why this subsection about transparency? The content is ok but I don't feel the need. With not about accountability and fairness as mention in line 430?	Remove it	
Ph.D.	493-497	6.7	§3	ed	Is it pertinent in this kind of document to develop concepts with instances?	Remove it	
Ph.D.	507-522	6.8	§2-§3	ge	Same remark as above. Here, the concept is quasi integrally explain by instances.	Develop and remove instances	
L.H.	529-530	7.1	§2	ed	If you use “e.g.” for example in one case you should use it in the other too.	Either remove the “e.g.” or add one: (e.g., a developer)	
L.H.	550-551	7.1	§7	ge	Redundancy with 525-526 (exact same line) Furthermore, it misses a transition for the next sections.	For the transition: To further illustrate the variety of expectations depending on the stakeholder, the next sections will focus on different actors for different scenarios.	
A.P.	552-558	7.1	Figure 1 & 2	ge	Both figures are not used in the text and this question how useful it is to provide them.	Either remove them or reference and use them in the text.	
L.H.	555-558	7.1	Figure 2	ed	It would be appreciated to have a higher resolution figure. In addition, a “?” has been forgotten in the title	Get a figure with better resolution	
L.H.	572	7.3		ed	Indent mistake	Remove indent	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
L.H. Ph.D.	573-575	7.3		te	For me those 2 scenarios fall into the category of “evaluation” methods described in Clause 6.6 I think that these 2 scenarios are for AI certifier profile.	Change the scenarios Add the AI certifier profile	
L.H.	579	7.3		ed	Replace “for AI developer is to input” with “for AI developer to contribute”		
L.H.	583-594	7.4		ge	In my opinion, it misses here an important scenario. For an AI product or service provider one of the key goals is to ensure its product does not discriminate an end-user because of its gender or its race or any attributes that will is legally punishable.	Add such a scenario	
L.H.	601-609	7.8		ge	In my opinion, it misses here an important scenario. Indeed, an evaluator will probably want to ensure that the system is not racist or sexist for example	Add such a scenario	
Ph.D.	559-630	7.2-7.13		ed	Is it pertinent in this kind of document to mention such example of scenarios?		
L.H.	647-650	8.3.1		te	As it is presented here it seems that one group has somehow the intent of using persuasive explanations for their own interest. Explanations could also be persuasive on their own because it plays with confirmation bias.	Add a subsection on confirmation bias (the explanations show what the AI developer wanted to see)	
L.H.	691-694	8.4.2		ge	It is likely that such standards also exist in aeronautics. It would be interesting to check if there are not works concerning pilots cognitive overload		
A.P. Ph.D.	631-701	8.		ge	A subsection on human bias is necessary. Indeed, many things could be wrong between an explanation and its understanding and thus its interpretation. There is confirmation bias that leads people to interpret what they expect. There	Add an 8.6 section “Consideration of humans cognitive biases” With subsections: - Confirmation bias	

1 **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

2 **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/NC ¹	Line number (e.g. 17)	Clause/Subclause (e.g. 3.1)	Paragraph/Figure/Table/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
					<p>is over-interpretation where humans extrapolate more than what they should. Humans abhor a vacuum and do not like to not be able to conclude, so they tend to invent something. Finally, there are misinterpretations where humans interpret something that is not consistent with the model's decision-making process.</p> <p>Example: a classification task, an image with a car and the model recognizes it. To explain this, we use attribution methods to explain it:</p> <ul style="list-style-type: none"> - <u>Confirmation bias</u>: The explanation covers all the car and some parts are highlighted more than the rest, such as the bottom and the front. We know that a car has wheels, license plate, grid for the motor... With such explanation, one could conclude that the model recognizes such elements. However, there is no proof of it; it only aligns with what was expected. - <u>Over-interpretation</u>: It can be close to confirmation bias, but here, we do not have prior expectations that influence our interpretation. We just conclude more than what we could. For example, if the model highlights the driver, one could say the model recognize that there is a driver holding the steering wheel. In fact, we have no idea. - <u>Abhor a vacuum</u>: When presented with an explanation, humans want to understand. It is easier to get a wrong understanding than admit we cannot conclude. The two previous elements are example of that. In fact, such cognitive biases are entangled. - <u>Misinterpretation</u>: The explanation was not precise enough and we interpreted 	<ul style="list-style-type: none"> - Over-interpretation - Abhor a vacuum - Misinterpretation 	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
					<p>something else. For example, the explanation highlights the wheel. We conclude that the model recognized the tire, but it could be the hubcap.</p> <p>Furthermore, the document should include ways to prevent them.</p>		
A.P. Ph.D.	631-701	8.		ge	<p>A subsection about “explanation communication” is necessary.</p> <p>Indeed, the visualization of an explanation can have a huge impact on the resulting comprehension and its interpretation. The communication manipulates the information for it to be intelligible for the stakeholder. However, this manipulation could remove, add, or change the initial information contained into explanation.</p> <p>This is not trivial and should be taken into consideration.</p>	Add an 8.7 section “Considerations on explanation communication”.	
Ph.D.	702-1590	9,10,11	all	ed	Why the term “explainability” in quotation marks?	Replace “explainability” with explainability	
Ph.D.	725	9.2	§3	ed	<p>“These properties can be...”. “can be” is very light and denotes a poor interest in Clause 10.</p> <p>What is the objective of this document?</p>	<p>Change for:</p> <p>These properties have to be formalized. The taxonomy described in Clause 10.2 provides a classification of properties applicable to explainability methods and approaches discussed in paragraphs bellow. This taxonomy can be used to harmonize AI system developments.</p>	
L.H.	746-750	9.3.2	§2	ed	As the paragraph is written, it seems that a user should select a single XAI method. In practice, you should identify a set of applicable XAI methods and compare or leverage all of them.		
Ph.D.	751	9.3.2	§3	te	Cf. Interpretation vs explanation discussion		

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
					See comments bellow (page 15) about clause 10.4.4		
L.H.	751-758	9.3.2	§3	te	It is not because you have a model explainable-by-design that it should prevent confronting the explanations yield with a post-hoc XAI method. In contrary, it would be a safe line to use several approaches and to compare and oppose them.	Add a sentence that says that nothing prevents using different approaches even on model explainable by design	
Ph.D.	771	9.4.1		ed	There is a lot of 'ty' here! I am not sure that the sentence denote the purpose. We consider the AI system V&V, not the explainability method V&V.	Remove at least "capability".	
Ph.D.	777-786	9.4.2	Note	ge	I think that some main characteristics have (or not!) to be listed. However, the complete list is difficult to establish and recommendations difficult to be provide. The list can be enrich with "Fairness", "Transparency", "Completeness of explainability", "Precision of explainability", etc. All these terms contribute to XAI evaluation. Depending the AI system criticality, stakeholders, human audience, model embedding, operational constraints (ODD), etc. the characteristics have to be prioritize and evaluated according these priorities.		
L.H. Ph.D.	788	9.4.2.1	§1	ge	The term "saliency maps" has not been defined yet. Saliency maps are just examples of kinds of results	Remove the first sentence	
Ph.D.	789-791	9.4.2.1	all	ge	Remark: Explainability methods are generally brings into software library with metrics algorithms. These metrics provide such	Add paragraph about explainability metrics	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
					measurement in order to score explainers. I think that this concept should appear more earlier in the document.		
Ph.D.	797	9.4.2.2	§1	ge	It is better to use the term Fidelity than Faithfulness. Faithfulness is about relationship (a man and his wife); Fidelity is about report (a movie adapted from a book). Even if these terms are confused in the scientific community, it is preferable to use only one.	Replace Faithfulness with Fidelity in the entire document.	
L.H.	805	9.4.2.3		ge	Is it really the conciseness that one is interested in? For example, saliency maps are not concise at all (224x224 pixels is a lot). A reflexion should be made here as it is more the ability to read the explanations that matters than its size.		
Ph.D.	813	9.4.2.4		ge	I don't understand "alignment with explanation needs" subsection in "evaluation of eXplainable Artificial Intelligence"? The alignment with explanation needs are validated upstream during AI inception (clause 9.2).		
Ph.D.	815	9.4.2.5		ge	What does it means?		
L.H.	817	9.4.2.6	§1	te	In addition to be robust to minor modification it should be deterministic i.e. for the same input my explanations should remain the same	Start with: In addition to be stable, that is deterministic, explanations should be robust...	
L.H.	878-879	9.7		ed	Mistake	Replace "re-evaluation stage" with "continuous validation stage"	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
Ph.D.	905	10.2.1		ge	The interpretation is no a need. It is a consequence of understanding the AI system.	Remove “interpretation or” or replace “interpretation or explanation” with “explanation and understanding”	
Ph.D.	925	10.2.2	§4	ed	Dot after domain.	domain. In the case...	
Ph.D.	930-931	10.2.3		ge	Cf. Interpretation vs explanation discussion	Remove “interpretation or” or replace “interpretation or explanation” with “explanation and understanding”	
Ph.D.	959	10.2.4		ge	Cf. Interpretation vs explanation discussion Even more here, it cannot be possible to characterise interpretation as local or global.	Remove “or interpreted” or replace “or interpreted” with “and understood”	
Ph.D.	962-963	10.2.4	§2	ge	Global interpretation...	Remove the last sentence or replace “Global interpretation” with “Global understanding”	
Ph.D.	965-966	10.2.4	§2	ge	Local interpretation...	Remove the last sentence or replace “Local interpretation” with “Local understanding”	
Ph.D.	979	10.2.6	§1	ge	Cf. Interpretation vs explanation discussion	Remove “interpretation or” or replace “interpretation or explanation” with “explanation and understanding”	
Ph.D.	992,101 5	10.2.7	Title, §7	ge	Fidelity vs Faithfulness	Replace Faithfulness with Fidelity	
Ph.D.	993,997, 1001,10 11,1015, 1017	10.2.7	§1,§2,§3,§6, §7,§8	ge	Fidelity vs Faithfulness	Replace system-faithful with system-fidelity	
L.H. Ph.D.	1010	10.2.7	§5	te	In addition to be typically hard to relate with the inner workings it is also prone to confirmation bias and be totally ignoring the system itself.	Add: It is typically hard to relate with the inner workings of the system. In	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
					e.g.: Looking at the closest element of a query in the input space. It is basically human to consider that similar element should receive similar treatments, however that might be irrelevant of what the system is doing as that kind of explanation (nearest neighbour) in the input space does not even consider the system!	addition, they may be infidel, that is they do not reflect at all the system's decision process and prone to confirmation bias, i.e. so human-aligned that it reveals what the users wanted to see.	
L.H.	1011	10.2.7	§6	ed	It would help for understanding to add an example here.		
Ph.D.	1011	10.2.7	§6	te	Such methods of explanation do not explain models nor AI systems. They extrapolate the decision-making process. It is important to note this by excluding them from the scope of the document	Add: Such methods of explanation do not explain the AI system. They extrapolate the decision-making process. They cannot be retained as tools for explainability of ML models and AI systems, the scope of this document.	
L.H.	1060	10.3.3		te	It would be nice to specify the kind of audience that those explanations target as it is done for the previous section		
L.H.	1069-1072	10.3.5		te	It should be mentioned that the size of the tree is significant to be relevant in a comparison with human decision process.	Modify the final sentence: Decision-tree explanations are typically useful for comparison with human decision processes, thus its depth is particularly significant as deeper tree will be harder to challenge with human reasoning.	
L.H.	1116-1118	10.4.2	§5	te	It is a little bit surprising that it would requires "AI-specific" skills. Indeed, as described here a statistician with the ability to query the model could do such a work without any AI-specific skills.	Remove the first sentence and the "However,"	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
Ph.D.	1120,1124	10.4.3	Title,§2	ge	Cf. Interpretation vs explanation discussion	Replace “Post-hoc interpretation” with “Post-hoc explanation”	
L.H.	1120	10.4.3	§1	ed	Typo mistake	“explainability” -> “Explainability” so it matches previous and incoming paragraphs	
Ph.D.	1126,127,1130,1132,1136,1138	10.4.4	Title,§1,§2,§3,§4	ge	Cf. Interpretation vs explanation discussion	Replace “inherently interpretable” with “inherently explainable”	
L.H.	1130	10.4.4	§2	te	It would be better to say inherently explainable, as decision trees are not inherently interpretable. A linear regression with thousands of parameters is very unlikely to be interpretable	Change “inherently interpretable” to “inherently explainable”	
Ph.D.	1133	10.4.4	§2	ge	Cf. Interpretation vs explanation discussion	Replace “degree of interpretability” with “degree of explainability”	
L.H.	1153	10.4.5	§5	te	There is no definition of symbolic AI and subsymbolic AI	Add those definitions in the clause 3	
Ph.D.	1161	10.4.5	§6	ge	Fidelity vs Faithfulness	Replace faithfulness with fidelity	
Ph.D.	1183	10.5.3	§1	ed	Typo mistake	“explainability” -> “Explainability”	
L.H.	1232	11.3.1.1	Table 2	te	Saying that all surrogate models do not have system-faithfulness is wrong or at least it is a harsh conclusion. Indeed, it depends on the surrogate model and how it was designed.	Remove the “No system-faithfulness”	
Ph.D.	1208-1590	11	all	te	The title of this section is “Approaches and methods to explainability” but the content is focused on methods even algorithms. It is interesting, but it will only require a large survey that is valid on a fixed date.	Develop approaches rather than methods without mention algorithms.	
Ph.D.	1208-1590	11	all	te	As possible, add the references of papers for methods and algorithms mentioned.	Add references	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/ NC ¹	Line number (e.g. 17)	Clause/ Subclause (e.g. 3.1)	Paragraph/ Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
L.H	1208-1590	11		ge	<p>While this entire section is pleasant to read and is an impressive survey, it probably does not belong in this document.</p> <p>First, the required technical level is much greater in this section than in the previous ones. Thus, it is probably not targeting the same audience.</p> <p>Secondly, it is neither an exhaustive listing nor a short overview of the existing XAI methods. However, there is no rationale on how and why some methods are in this document.</p> <p>Third, it is not clear as presented here that a given method could have been introduced or described in other clause (ex: LIME belong to both surrogate model & post-hoc methods).</p> <p>While I am personally interested on having such a classification of all XAI methods with properties of interest, it is, in my opinion, out-of-scope for this document.</p>	<p>Describe more generally the families of approaches introduced here with global properties of a family. Points out the strengths and the flaws of each of them (possibly with nuances). Examples could be given but reference them rather than describing it thoroughly here.</p> <p>Do not forget in the first section that some methods can belong to one or several families of methods.</p>	
L.H.	1591-1606	Annex A		ge	It is interesting but it would be much better if some parts (other sections) mentioned and pointed them.		
L.H.	1673-1676	B.2.3		te	Description of activation atlas is vague thus; it is not easy to understand it.		
L.H.	1703-1717	B.2.5		ge	This example is striking but it seems also unrealistic. Indeed, it is very unlikely to have a model using those 2 features and to come up with such a complex reasoning. The idea that some features correlate to exogenous ones is good but it is hard to buy that ones would be able to generate explanations with exogenous features.	Maybe saying that there is a high correlation between people eating anchovies and thus having a disease (so it is not an exogenous feature anymore) especially when the patients do not eat other fishes	
L.H.	1765-1813	B.2.6		ge	It would be better to separate faithfulness and human-alignment to avoid confusion. Those two	Split in two sections: Faithfulness and Human-alignment without mixing the two.	

¹ **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

² **Type of comment:** **ge** = general **te** = technical **ed** = editorial

Template for comments and secretariat observations

Date: 19/07/2023	Document: ISO/IEC TS 6254:2023(E)	Project: ISO/IEC JTC 1/SC 42/WG3
------------------	------------------------------------------	----------------------------------

MB/NC ¹	Line number (e.g. 17)	Clause/Subclause (e.g. 3.1)	Paragraph/Figure/ Table/ (e.g. Table 1)	Type of comment ²	Comments	Proposed change	Observations of the secretariat
					are clearly different so it should appear in the document.		

1 **MB** = Member body / **NC** = National Committee (enter the ISO 3166 two-letter country code, e.g. CN for China; comments from the ISO/CS editing unit are identified by **)

2 **Type of comment:** **ge** = general **te** = technical **ed** = editorial