Accessible Cultural Heritage through Explainable Artificial Intelligence

Natalia Díaz-Rodríguez, Galena Pisoni

Accessible Cultural Heritage through Explainable Artificial Intelligence

Natalia Díaz-Rodríguez

natalia.diaz@ensta-paris.fr ENSTA, Institut Polytechnique Paris & INRIA Flowers, France

Galena Pisoni

galena.pisoni@univ-cotedazur.fr, Université Côte d'Azur, Nice, France

Abstract

Ethics Guidelines for Trustworthy AI advocate for AI technology that is, among other things, more inclusive. Explainable AI (XAI) aims at making state of the art opaque models more transparent, and defends AI-based outcomes endorsed with a *rationale explanation*, i.e., an explanation that has as target the non-technical users. XAI and Responsible AI principles defend the fact that the audience expertise should be included in the evaluation of explainable AI systems. However, AI has not yet reached all public and audiences, some of which may need it the most. One example of domain where accessibility has not much been influenced by the latest AI advances is cultural heritage. We propose including minorities as special user and evaluator of the latest XAI techniques. In order to define catalytic scenarios for collaboration and improved user experience, we pose some challenges and research questions yet to address by the latest AI models likely to be involved in such synergy.

Keywords: Explainable Artificial Intelligence, Generative Models, Natural Language Processing, Image Captioning, Cultural Heritage

1. Introduction

The European Commission Ethics Guidelines for Trustworthy Artificial Intelligence (AI) [1] and Responsible AI principles [2] advocate for lawful

Preprint submitted to PATCH Workshop at the 28th ACM UMAP Conference, May 26, 2020

Figure 1: Left: *3 Graces*. Middle: *Monet* from the series *People matching artworks*. Right: *People touching artworks*. Reproduced with permission from ©Stefan Draschan `www.StefanDraschan.com`.

AI technology that is, among other things, more inclusive. *EXplainable* AI (XAI) aims at making state of the art opaque models more transparent, and defends AI-based outcomes endorsed with a *rationale explanation*, i.e., an explanation that has as target the non-technical users. The latest XAI techniques [2, 3, 4, 5] could bring art closer to new audiences. By increasing the accessibility of cultural heritage to collectives not fully able to enjoy it today, missing gaps in technology could be identified. One example of such innovations is the smartphone app MonuMAI[1], which has already demonstrated how to put together technological innovation to actively approach different perspectives in science and art dissemination to the public [6, 7]. Based on deep neural networks (DNNs), MonuMAI classifies photos taken (e.g. of a facade) according to different architectonic styles, providing visual explanations on the elements contributing to the detected style.

Such examples show that technology can yet have a lot more of impact than currently has. Models able to switch among input/output modalities (in terms of the data they are able to process) could have a crucial role. The role is actively approaching art to minorities not having it accessible (since blind people can listen and read, the deaf can read, etc.). The latest advances in natural language processing (NLP), computer vision (CV) and XAI could disruptively innovate the ways in which we teach, learn, and approach art to

[1]MonuMAI = Monuments + Maths + AI + Dissemination

society.

For instance, people with visual impairments take and share photographs for the same reasons that sighted people do, but as they find many more difficulties, methods have been developed to assist blind photography (including audio feedback that facilitates aiming the camera) [8]. Generating descriptions helps visually impaired people better browse and select photos based on human-powered photo descriptions and computer-generated photo descriptions. Could such human computation-generated visual explanations also help completely blind users, e.g. to navigate? Could these help any user that wants to learn from first-hand experts how a given artwork is interpreted, or what it conveys, providing the right context of its time? If the answer is positive, perhaps a DNN could be trained with all generated data to avoid the arduous task of labelling data so that eventually, the blind would not require human assistance. In this paper we put ourselves in the shoes of particular collectives such as the blind, or the deaf, and pose a set of settings we consider worth exploring in the intersection of art and science. In particular, we propose using cultural heritage as a playground for (X)AI, and suggest a list of challenges and research questions (RQs) showing why inclusive art needs XAI, and why XAI may find on minority audiences, the right manner to evaluate where AI can have more impact.

2. EXplainable AI (XAI)

Given an audience, an ***eXplainable AI*** *(XAI) is a suite of machine learning techniques that produces details or reasons to make its functioning clear or easy to understand* [2]. XAI draws insights from Social Sciences and the psychology of explanation, and its objective is to (1) produce more explainable models maintaining high level performance, and (2) enable humans to understand, trust, and manage the emerging generation of artificially intelligent partners.

Given the inherent subjectivity of an explanation, current discussions advocate for rethinking interpretability, involving the audience expertise. When AI becomes ubiquitous across domains, it is specially important to follow the EU Ethics Guidelines for Trustworthy AI [1], Guidelines for Responsible AI and interpretable AI models [2]. Equally important is accounting for interests, demands and requirements of the different stakeholders interacting with the system to be explained. In cultural heritage contexts, accounting for the target audience is equally important from both evaluation and personalisation points of view [9].

3. Unconventional interfaces for art accessibility

Groups of visitors inside museums have been a focus of ongoing research for a long time [10, 11, 12]. Some systems allowed for visitor collaboration by supporting shared listening or leaving messages between visitors [11]. In order to facilitate the process of engagement and collaboration between co-visitors, narratives are often introduced in museum contexts. Narratives are responsible for mental immersion through which users can be engaged and involved in the experience, increasing their sense of mediated presence as well. Visitors preferences have been studied [13], and more engaging approaches have been proposed for stimulating the visitor interests by using presentations such as film or drama [14]. The drama was adapting to the visitors so that different available independent drama segments were played to be group based on characteristics of the group of visitors, the specific context of the visit, and implicit input from the visitors themselves. Results showed that drama, when designed for small groups, and combined with the raw emotion of onsite visitors being in front of actual original artifacts, can emotionally engage distant visitors with mobility constrains [15].

Another way to alleviate mobility disadvantages for challenged individuals and to allow them still to enjoy art is through the use of virtual environments. Virtual environments offer the possibility to navigate in new or known environments and contexts, and interacting with people in different locations. Virtual environments can provide a realistic experience, or the participant's feeling of "being there" in an environment, also defined as a sense of presence. Previous studies have investigated if and how challenged individuals can access and appreciate museum contents, and the best suited interface designs for this [16, 17]. The results have been positive with first results indicating that challenged individuals could indeed understand the virtual tours and engage in contextual conversations, while the ability to follow the tour depended on the level of the "interactivity" of the prototype. The more complex the interaction, the least possible it was for challenged individuals to follow the museum visit.

For those with cognitive disabilities and the elderly, the ability to consume cultural contents and to independently consult information about museums from home is even more limited. Previous applications that understand the cognitive barriers and propose solutions to present information so to cope with the reduced cognitive loads have been developed and tested with users [16, 18]. The majority of studies focus on developing or using AR technologies to support blind or visually impaired users. Successful steps towards this future have been made, with the possibilities for shared experiences already available also for people with cognitive disabilities.

3.1. Storytelling and audience engagement

Approaching art to different audiences should consider culture and background. Culture traditions can disruptively change the idea of a museum activity since early ages. For instance, opinion towards museums can be seen by kids very differently. A great example is how kids loudly enjoy and see museums as a fun place for kids when allowed to paint and talk inside (as in UK National Gallery). The idea of museum becomes that one of a ludic place, transmitting the idea that art can be a fun activity to play with. Such context makes kids at ease to approach and feel curious about heritage, leaving room for creativity. A very different idea of art is what often is formed in children when museums do not allow touching, loud speaking, nor interaction, linking the idea of museum more to a temple, or an activity that many may find boring.

Studying mechanisms to bring closer the artistic heritage to a target audience shows that, in art, the audience plays a central element, and can change the vision of society towards art dramatically. Likewise in XAI, not placing the audience in a centric role risks AI losing its deserved trust.

In order to renew the ways of thinking about art,

Challenge 1. *Could AI help deliver art, personalize or write new rules on what is possible to do with cultural heritage?*

Neural symbolic computation [19, 20] includes methods to embed symbolic and neural representations to learn and reason with different levels of abstraction [21].

RQ 1. *Does embedding of expert/domain knowledge into DL models [22] help explain such models? Can XAI help encode such prior knowledge [22]?*

Use Case 1. *Juan Jesus Pleguezuelos, History teacher and podcast author of Art History for entrance exams to University[2]: The challenge I pose is to*

[2] https://www.instagram.com/historiaarte.selectividad

make others see an historical image only through words. It is clear that this requires an exhaustive description of the masterpiece, but you should also try to make others feel the latent soul in it, and decipher the intention of the author. And if you could also convey the emotion that this work is able to cause, it can be that words may be more than enough to make a listener understand an artistic work that he is not seeing in that moment.

Challenge 2. *Could XAI exhibit the level of detail and engagement required to effectively convey a style, or the spirit represented in the times of an artwork?*

4. Explaining art through language

Unlike math, art may not always be understood, and may require extra (objective and subjective) interpretations to be able to effectively convey its message. We believe art and the story accompanying it could be made more widely understood if they would be more easily accessible.

Hypothesis 1. *If AI models can assist generation of content- and interpretation-wise explanations, art can be more widely understood and accessible.*

One difficulty to convey the style of art eras consists of the ability to express what that era meant. E.g., Renascence's works show people's joy, elegance, etc. AI not only should recognize the style but also the spirit present in the era. For instance, given *Venus Birth*, how is to be understood the Renaissance period? How to understand the ideas and spirit of the time? What was the intention of the author? XAI may be a well-fitting candidate tool to help this objective, being a catalyst for on-demand interfaces to truly adapt to every active audience.

Producing textual explanations through NLP is a way of explaining AI models [2]. Image captioning, visual question answering (VQA) and textual advisable explanations are different ML tasks considered. An example of advisable explanations is on computer vision scene understanding for autonomous driving learning models [23].

4.1. Image captioning models

Image captioning models produce a text describing the scene given an input image. With the aim of producing clarifying explanations on why a particular image caption model fails or succeeds, since a deep neural network (DNN) is considered a black box model hard to inspect, recent strategies make sure that the objects the captions talk about are indeed detected in the images [24, 25]. Textual explanations can also contribute to make vision and language models more robust, in the sense of being more semantically grounded [26].

Since image captioning models pretrained on datasets outside the art domain fail completely at describing out of distribution inputs (e.g., pictorial compositions not found in natural images), some metrics evaluating the semantic fidelity of the model have been devised [24]. These call for models more semantically faithful to the input information, in order to reduce the bias that image caption models suffer [27], as well as object hallucination. The latter is a well-known phenomenon where image captioning models caption an image with objects not present in the image [28].

Captioning models including sentiment have also been developed [29], either using the viewer's attitude and emotions towards the image [30], or including emotional content inherent to the artwork image [31].

Hypothesis 2. *(X)AI can explain art.*

Content vs Form

RQ 2. *Could (X)AI distinguish among a) content vs b) form explanations? Could (X)AI produce a) content and b) form explanations?*

The above RQs highlight the challenge of synthesizing figurative sense (interpretation) vs literary sense (content) explanations of an artwork.

4.2. Visual Question Answering models

Another NLP model to produce explanations about an image is tackled by the problem of visual question answering [32], specially useful for the blind[3] or image captioning projects[4][5][6]. Generating questions that can be answered

[3] https://vizwiz.org/

[4] lens.google.com Google Lens is an *image recognition technology designed to bring up relevant information related to objects it identifies using visual analysis based on a neural network*.

[5] Google Goggles was an image recognition mobile app used for searches based on pictures taken by handheld devices.

[6] https://lazarilloproject.github.io/

by a DNN's output caption can improve explainability and quality of image captioning models [33].

RQ 3. *Could art explanations be generated on request, i.e., using visual question answering (VQA)?*

Advisable text explanations have shown to be useful when teaching models to drive autonomously [23].

RQ 4. *Could advisable explanations increase the engagement and interest in artwork?*

To enrich the experience of a user when observing art, an advisable interactive introspection explanation could be: *Pay attention to where the light is set in this painting. What is the center of focus the author is highlighting as such? Why?*

RQ 5. *Should only objective or also subjective information be part of an artistic explanation?*

5. Explaining visual art through generative and multimodal models

Generative adversarial networks (GANs) are considered a form of artificial curiosity [34]. Generative models have been successfully used for image inpainting [35, 36] or image reconstruction. A potential application of inpainting, i.e., filling the gaps in a given image, could be 2D or 3D restoration [37]. For instance, DAFNE (Digital Anastylosis of Frescoes challeNgE) dataset[7][38] allows to design methods to aid conservators and restorers perform fresco reconstruction when pieces are missing, spurious or suffer erosion.

Another application of generative models is performing style transfer. Style transfer models successfully disentangle the data generating factors [39] such as content and style when synthesizing paintings [40]. Similarly, music instruments can be extracted from videos [41] using multimodal CNNs.

RQ 6. *Can XAI disentangle the underlying data generating (historical, stylistic, spiritual) factors behind a generative model output?*

[7]It considered inclusion of autism users `https://vision.unipv.it/DAFchallenge/DAFNE_dataset/`.

Edmond Bellamy (Fig. 2) was the first piece of AI (GAN)-generated art to come to auction at Christie's, demonstrating that algorithms are able to emulate creativity[8].

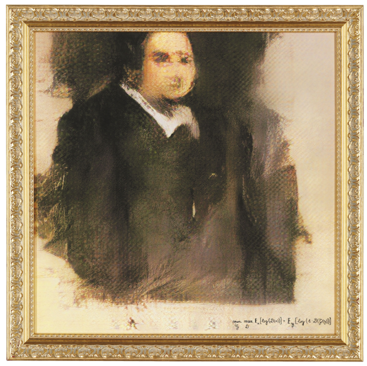

Figure 2: *Edmond de Belamy*. Credit: ©Obvious, 2018 (instagram: @obvious_art)

Explainable AI techniques could assist explaining what artists and styles influenced themodel training the most, in order to apply feature attribution methods to rate most prominent influence, helping perhaps understanding what elements made it succeed.

Challenge 3. *Can XAI explain a given artwork' success in terms of the underlying influencing artistic styles?*

For instance, what makes disruptional and interesting Trina Mery artistic body painting compositions[9], Stefan Draschan's photography, or Prof. Pleguezuelos's History podcasts[10], or *Edmond Belamy*?

Dreaming machines using multimodal data fusion and information retrieval are an example of neural-symbolic cognitive agent that can hallucinate visual input when it is completely or partially blanked (mimicking loss of vision) [43].

RQ 7. *Could models learn to hallucinate a missing data modality given a lack of the privileged information [44]?*

[8]Sold at $432,500 [42] `https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx`

[9]`https://www.trinamerry.com/`

[10]`https://www.instagram.com/elprofesorinquieto`

Biologically plausible models such as Deep Boltzman Machines' sensory hallucinations could be generalized to potentially validate the understanding of a deep neural network (DNN) and verify whether its output is faithful to the original content of the artwork. Perhaps in the same manner a machine can learn to explain non regular input modalities, e.g. touch-based artwork, through words or sounds.

6. Art and Robotics

Creativity is consider a driver for research in robotics in open ended learning environments [45], because performance is not the only criteria to be assessed on robots when they must learn to deal with new situations. In these cases, creativity can quantitatively measure progress, define diversity-driven behaviours, or deal with unforeseen damages [46].

In terms of accessibility, technological advancements have brought "telepresence" or mobile remote presence (MRP) systems as another opportunity for bridging social and spatial barriers for people with mobility constrains. MRPs are designed to be teleoperated and are used to improve communication between individuals. They were found to have the potential to assist challenged individuals in instrumental activities of daily living as well as to foster social interaction between people. A number of qualitative studies where people with mobility constrains used an MRP system identified benefits for the participants such as being able to see and to be seen, reducing costs and hassles associated with traveling, and reducing social isolation [47]. Experiences with an interactive museum tour-guide robot have been described in previous literature [48]. Questions on how to provide the same user experience, while users teleoperate a robot to make the experience as close as possible as if they were there physically are still to be solved.

Learning joint representation models from vision and language is useful for navigation of embodied robotics [49]. On the other hand, robotics can be thought of as delivery means for art explanations. For instance, a robot can sense when the group he is leading in Seville's Alcazaba tour is getting bored, and change, e.g., the length of its explanations based on the movement of the visitors [50, 51]. In this context, it is worth investigating the utility of such robots in terms of:

RQ 8. *Do remotely operated mobile robots increase virtual visits to a cultural site, with respect to static browser-based virtual tours?*

RQ 9. *Do robot guides [50] improve the visitors rating when no human guide is available? Is their user experience rated better than walkytalky guides?*

RQ 10. *Can AI provide guide explanations that reduce the boredom of the visitors?*

There could be a value in having a AI-empowered robots visiting together the cultural heritage site with the humans as well. One potential application and advantage of using robots and AI in cultural heritage is with respect to language: e.g., a robot like C-3PO that speaks all languages can make the tour anytime in any language, including sign language. This has a value with respect to a human tour guide and can be seen as a next step in innovation in the field of guide systems, as the incarnation of audio guides.

Other types of robots have created art on their own. A Russian research group developed a robot which incorporates a novel colour-mixing device that can, in principle, create any shade or hue. The researchers used both off-the-shelf components and 3D-printed parts to build their robot. It includes an algorithm that transforms a photographic image into a set of vectors that programs the robot's brush to imitate human brushstrokes [52].

7. A call for a multidisciplinar collaboration

The presented challenges aim at stimulating a call for collaborators in a joint effort to mutually learn from other domains, and form an interdisciplinary research consortium aggregating a diverse set of collective and symbiotic needs:

- Art historians: can gain visibility by making art accessible, building a portfolio, e.g. as gallery guides, art podcast content generators, etc. Humanities students could better learn by teaching their lessons outside humanities and generating AI-consumable data.

- Artists and story tellers could earn an audience willing to learn about a niche passion.

- Disabled and minorities: The blind could get access to art explanations through audio or text resources, the deaf through the latter's transcriptions.

- Computer scientists would use the generated data to build robust machine learning models that (1) explain art, and (2), are explainable.

The ultimate aim is that all content would facilitate anyone to understand any art with the right context.

7.1. Impact of AI on Technological Domains

We envision a set of domain areas where the symbiosis among art and (X)AI could be further exploited. In order to guarantee Responsible AI guidelines [2], provenance specification of XAI training and generated resources should be a requirement.

Recommendation systems and personalization services may optimize matching art-tellers and art-listeners, and suggest new artworks likely to be appreciated by a given public.

Educators and developmental psychologists could find in XAI a support tool to convey humanities, social sciences and history in terms of the alignment of explanation facts with the mental model and cultural background of the learner.

After all these technologies are put into place, and human in the loop machine learning systems have gathered enough data, a new wave of creative AI algorithms will emerge. All byside data generated through Human-Machine collaborations involving the stakeholders above could train deep models to capture the underlying generating factors that make humans interpret art the way they do.

However, language could perhaps transfer art across domains, adapting accordingly to the requested format and medium at each time.

Because language cannot express art, but is the closest mean for universal communication, we expect art expression through deep and word-based representations to be one form of universal intermediate language allowing to sing a painting, or to draw a song.

Challenge 4. *Tackling the lack of personal touch in technology*

During quarantine/crises, diverse cultural agendas are made available for free (operas, museums, virtual tours, circus, libraries, etc.). At-home vs onsite experiences can degrade the experience of culture, perhaps due to lacking the social touch involved in the original experience. Human computation, art-history and humanities expertise on the approach to such cultural offer could not only serve the purpose of bringing art home, but also set the basis for future ML models that could generate personalized explanations about a given artwork. A hypothesis is that museum experiences require of a personalized, social or physical involvement experience in order to maximize the inherent pleasure of enjoying cultural heritage sites, with everything that it conveys.

ML algorithms generate sketches [53], steerable playlists [54], music [55], and incite creativity through editing tools [56]. Since machine discriminators outperform humans in detecting generated text [57],

RQ 11. *Could AI recognize XAI generated explanations better than humans?*

RQ 12. *Can human testimony personalized art explanations stimulate engagement and discovery of art by society?*

RQ 13. *Could artist voice note explanations uplift the lack of social touch in traditional virtual/ audio guides?*

We hypothesize:

Hypothesis 3. *Digitized artwork personal reviews can enrich access to cultural heritage based on artists audio/transcriptions, making it available to any art consumer, including the deaf and the blind.*

Challenge 5. *Evaluating AI-generated art explanations*

RQ 14. *Is XAI being evaluated in the right tasks and with the right audience?*

RQ 15. *Can we evaluate AI generated text explanations' quality in a quantitative manner that is both user questionnaire-free and audience-specific?*

In order to assess story quality, word embeddings can be used to estimate cognitive interest [58, 59, 60]. Fashion styles and its social media tags can be used to predict subjective influence and novelty [61]. Could such influence and novelty metrics correlate with actionable or useful explanations?

RQ 16. *Could AI explain what makes an artwork appreciated or liked? Could we quantify the amount of surprise or originality it conveys?*

Challenge 6. *Defining explanation standards*

RQ 17. *Can we define standards for XAI explanations, including those subject to subjectivity?*

General XAI techniques usually evaluate XAI techniques on their ability to generate visual or textual explanations [2]. However, the requirements to evaluate an explanation positively by a blind or deaf person are likely to require very different criteria.

RQ 18. *Can we always provide automatic satisfying answers when the observer is unable to see/ visually impaired?*

Challenge 7. *Explaining concepts hard to visually grasp*

A single format may not fit to convey all art modalities. At times, some modalities, e.g., sound, may be a better format to translate into. However, visual-textual semantic embedding [62] and retrieval [63] is possible. In the latter case, without labelled cultural heritage data thanks to transfer learning.

If *what is essential is invisible to the eyes*[11], symbols such as words or knowledge graphs could act as intermediate proxy representation to verbalize complex abstract concepts.

RQ 19. *Can multimodal deep representations be an intermediate language to universally convey art? Could these generate text explanations for tech and non technical audiences?*

7.2. XAI as a medium, rather than a menace to human creativity

Historians can argue that humanities education can currently abuse the use of images to teach. This is demonstrated by the success of an influencing teacher's podcast that prepares for History university entrance exams. *While the use of words stimulates the imagination and keeps the mind working, providing an image to explain the same concept keeps the mind static and inactive.* This is why teacher Pleguezuelos points to the images corresponding to the podcast explanations in *Instagram*[12] only after students had to

[11] *It is only with the heart that one can see rightly; what is essential is invisible to the eye. -Antoine de Saint-Exupéry*

[12] `https://www.instagram.com/historiaarte.selectividad`

imagine the described period, era, or artwork, exclusively with words. Could a machine learn the same way? Could it reinforce the knowledge through later confirmation with a different learning modality?

Challenge 8. *What is the key role that AI can play in bringing heritage closer to the viewer?*

An artwork can inspire our mind if we are taught in what epoch it was represented, and in what context it was created. If AI models could ever be powerful enough to make us re-live that era, the inspiration they transmitted, and even imagine the spirit of the age,

Challenge 9. *Could AI destroy the creativity of the viewer, that part that inspires the audience?*

We argue that since AI can learn from a multimode of inputs, it can provide interesting analogies or links to other artworks that a human could not do. XAI techniques should explore ways in which AI could be not a threat to the development of creativity that the artwork itself implies, but rather a facilitation medium that suggests questions, allows exploring unknowns, and further stimulates scientific curiosity and hunger for knowledge. In this context, artificial models of computational curiosity [64] could align with those of humans, to guide the latter to improve its mental model, trust, and curiosity [65]. Curiosity increase could act as metric of positive understanding of art and its whole context.

8. Discussion and Conclusions

Panels discussing the abilities of computational creativity involving scientists and humanities can results in fiery discussion[13]. Research labs in Digital Humanities investigate perceptual and cognitive tasks related to human creativity. This shows that, as in developmental robotics where robot models are trained for open-ended learning [45], having to perform life-long learning [66] continually, both humans and machines can learn from each other, better inform hypotheses and experiments, and allow synergistic research.

[13] *Computational Creativity: Art through the Eyes of Computation* (panel arranged by N Díaz-Rodríguez & S Tomkins, Data Science Santa Cruz initiative, including art historians, computer scientists, musicians and humanists): `http://ihr.ucsc.edu/event/quantifying-creativity-art-through-the-eyes-of-computation/` Video: `http://travellingscholar.com/qcreativity/`

Challenge	Dimensions	Concerns
Augmenting accessibility to minorities or users with physical & cognitive disabilities	Interface and content personalisation, Generative and Multimodal AI	Inclusion, AI Ethics
Making AI explainable	Explainable & Interpretable AI	FAT ML, Responsible AI
Explaining art with AI	Human computation, Human in the loop, [Multimodal, Generative] AI	Trust, Responsible AI
Creativity as research engine, AI for content synthesis	Engagement, Curiosity, Computational creativity	Trust, Subjectivity Metrics

Table 1: (X)AI for Cultural Heritage Challenges.

We summarized hypothesis and RQs into challenges, discipline dimensions affected and concerns to address such challenges in Table 8. We presented some disruptive art settings as motivating examples where AI and XAI could have novel research playgrounds to validate models. Since concerns involve fairness, accountability, transparency (FAT) in ML, we gave a first step listing questions that need to be addressed to obtain insights on how AI can best help accessibility to audiovisuals.

Despite having presented here challenges and opportunities focused on how AI (and robotics) can help access cultural heritage and the digital humanities, this is just an application domain where the limits of current AI models can be stress-tested. The existing challenges to attain explainable AI in any real-life problem are equally relevant and should be explored, especially in practical applications of AI safety and AI for social good (from elderly telepresence robots [67] to epidemic and hospital crisis management [68]).

9. Acknowledgements

We thank Juan Jesús Pleguezuelos for the motivating testimonies, Stefan Draschan and OBVIOUS collective for borrowing illustrating images, and Pranav Agarwal, Siham Tabik, Francisco Herrera, Alberto Castillo Lamas, Mario Romero, Serena Ivaldi and Paris Cité Universitaire friends for inspiring brainstormings.

References

[1] High Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, Technical Report, European Commission, 2019.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion (2019). URL: http://www.sciencedirect.com/science/article/pii/S1566253519308103. doi:https://doi.org/10.1016/j.inffus.2019.12.012.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[4] D. Gunning, Explainable artificial intelligence (xAI), Technical Report, Defense Advanced Research Projects Agency (DARPA), 2017.

[5] S. T. Mueller, R. R. Hoffman, W. J. Clancey, A. Emrey, G. Klein, Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI, CoRR abs/1902.01876 (2019). URL: http://arxiv.org/abs/1902.01876. arXiv:1902.01876.

[6] F. Herrera, A. Martinez-Sevilla, S. Tabik, R. Montes, A. Castillo, T. C. Sánchez, J. P. Cruz, Competicion caepia-app: Monumai, una app para incrementar el valor social del patrimonio-arquitectonico andaluz (2018).

[7] F. Fernández Morales, J. Valderrama Ramos, S. Luque López, A. Mart/'inez Sevilla, J. Policarpo Cruz Cabrera, P. Alvito, Paseos Matemáticos por Granada: Un estudio entre arte, ciencia e historia, Editorial Universidad de Granada, 2017. URL: https://dialnet.unirioja.es/servlet/libro?codigo=701550. doi:10.1007/978-3-030-22327-4_13.

[8] Y. Zhao, S. Wu, L. Reynolds, S. Azenkot, The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments, CoRR abs/1805.01515 (2018). URL: http://arxiv.org/abs/1805.01515. arXiv:1805.01515.

[9] C. Rocchi, O. Stock, M. Zancanaro, M. Kruppa, A. Krüger, The museum visit: generating seamless personalized presentations on multiple devices, in: Proceedings of the 9th international conference on Intelligent user interfaces, 2004, pp. 316–318.

[10] O. Stock, M. Zancanaro, PEACH-Intelligent interfaces for museum visits, Springer Science & Business Media, 2007.

[11] T. Kuflik, O. Stock, M. Zancanaro, A. Gorfinkel, S. Jbara, S. Kats, J. Sheidin, N. Kashtan, A visitor's guide in an active museum: Presentations, communications, and reflection, Journal on Computing and Cultural Heritage (JOCCH) 3 (2011) 11.

[12] P. M. Aoki, R. E. Grinter, A. Hurst, M. H. Szymanski, J. D. Thornton, A. Woodruff, Sotto voce: exploring the interplay of conversation and

mobile audio spaces, in: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 2002, pp. 431–438.

[13] G. Kostoska, D. Fezzi, B. Valeri, M. Baez, F. Casati, S. Caliari, S. Tarter, Collecting memories of the museum experience, in: CHI'13 Extended Abstracts on Human Factors in Computing Systems, 2013, pp. 247–252.

[14] C. Callaway, O. Stock, E. Dekoven, Experiments with mobile drama in an instrumented museum for inducing conversation in small groups, ACM Trans. Interact. Intell. Syst. 4 (2014). URL: https://doi.org/10.1145/2584250. doi:10.1145/2584250.

[15] G. Pisoni, F. Daniel, F. Casati, C. Callaway, O. Stock, Interactive remote museum visits for older adults: an evaluation of feelings of presence, social closeness, engagement, and enjoyment in an social visit, in: 2019 IEEE International Symposium on Multimedia (ISM), IEEE, 2019, pp. 99–993.

[16] G. Kostoska, M. Baez, F. Daniel, F. Casati, Virtual, remote participation in museum visits by older adults: a feasibility study, in: 8th International Workshop on Personalized Access to Cultural Heritage (PATCH 2015), ACM IUI 2015, 2015, pp. 1–4.

[17] G. Kostoska, A. P. Vermeeren, J. Kort, C. Gullström, Video-mediated participation in virtual museum tours for older adults, in: 10th International Conference on Design & Emotion, 27-30 September 2016, Amsterdam, The Design & Emotion Society,, 2016.

[18] M. Gea, X. Alaman, P. Rodriguez, V. Rodriguez, Towards smart & inclusive society: building 3d immersive museum by children with cognitive disabilities, in: Proceedings of the EDULEARN16: 8th International Conference on Education and New Learning Technologies, Barcelona, Spain, 2016, pp. 4–6.

[19] T. R. Besold, A. d'Avila Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kuehnberger, L. C. Lamb, D. Lowd, P. Machado Vieira Lima, L. de Penning, G. Pinkas, H. Poon, G. Zaverucha, Neural-Symbolic Learning and Reasoning: A Survey and Interpretation, 2017. arXiv:1711.03902.

[20] G. Marra, F. Giannini, M. Diligenti, M. Gori, Integrating learning and reasoning with deep logic models, 2019. arXiv:1901.04195.

[21] A. Bennetot, J.-L. Laurent, R. Chatila, N. Díaz-Rodríguez, Towards explainable neural-symbolic visual reasoning, in: NeSy Workshop IJCAI 2019, Macau, China, 2019.

[22] M. Diligenti, S. Roychowdhury, M. Gori, Integrating prior knowledge into deep learning, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2017, pp. 920–923.

[23] J. Kim, A. Rohrbach, T. Darrell, J. Canny, Z. Akata, Textual explanations for self-driving vehicles, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 563–578.

[24] P. Agarwal, A. Betancourt, V. Panagiotou, N. Díaz-Rodríguez, Egoshots, an ego-vision life-logging dataset and semantic fidelity metric to evaluate diversity in image captioning models, in: Machine Learning in Real Life (ML-IRL) Workshop at the International Conference on Learning Representations (ICLR), 2020. URL: https://arxiv.org/abs/2003.11743.

[25] J. Lu, J. Yang, D. Batra, D. Parikh, Neural Baby Talk, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 7219–7228.

[26] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, D. Batra, D. Parikh, Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded, 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 2591–2600.

[27] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, A. Rohrbach, Women also snowboard: Overcoming bias in captioning models, in: European Conference on Computer Vision, Springer, 2018, pp. 793–811.

[28] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, K. Saenko, Object Hallucination in Image Captioning, CoRR abs/1809.02156 (2018). URL: http://arxiv.org/abs/1809.02156. arXiv:1809.02156.

[29] I. Hrga, M. Ivašić-Kos, Deep image captioning: An overview, in: 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2019, pp. 995–1000.

[30] A. P. Mathews, L. Xie, X. He, Senticap: Generating image descriptions with sentiments, in: Thirtieth AAAI conference on artificial intelligence, 2016.

[31] O. M. Nezami, M. Dras, P. Anderson, L. Hamey, Face-cap: Image captioning using facial expression analysis, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2018, pp. 226–240.

[32] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, J. P. Bigham, VizWiz Grand Challenge: Answering Visual Questions from Blind People, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 3608–3617. URL: `https://arxiv.org/abs/1802.08218`.

[33] J. Wu, Z. Hu, R. J. Mooney, Generating question relevant captions to aid visual question answering, arXiv preprint arXiv:1906.00513 (2019).

[34] J. Schmidhuber, Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991), Neural Networks (2020).

[35] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[36] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, Y. Akbari, Image inpainting: A review, Neural Processing Letters (????) 1–22.

[37] A. F. Abate, S. Barra, G. Galeotafiore, C. Díaz, E. Aura, M. Sánchez, X. Mas, E. Vendrell, An augmented reality mobile app for museums: Virtual restoration of a plate of glass, in: Euro-Mediterranean Conference, Springer, 2018, pp. 539–547.

[38] V. Cantoni, L. Lombardi, G. Mastrotisi, A. Setti, The DAFNE Project: Human and Machine Involvement, volume 99, Electronic Imaging & the Visual Arts: EVA 2019, Florence. Firenze University Press, 2019.

[39] A. Achille, S. Soatto, Emergence of invariance and disentanglement in deep representations, The Journal of Machine Learning Research 19 (2018) 1947–1980.

[40] L. A. Gatys, A. S. Ecker, M. Bethge, A neural algorithm of artistic style, arXiv preprint arXiv:1508.06576 (2015).

[41] O. Slizovskaia, E. Gómez, G. Haro, Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture, in: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, 2017, pp. 226–232.

[42] G. Vernier, H. Caselles-Dupré, P. Fautrel, Electric dreams of ukiyo: A series of japanese artworks created by an artificial intelligence, Patterns 1 (2020) 100026.

[43] L. de Penning, A. D. Garcez, J.-J. C. Meyer, Dreaming Machines: On multimodal fusion and information retrieval using neural-symbolic cognitive agents, in: A. V. Jones, N. Ng (Eds.), 2013 Imperial College Computing Student Workshop, volume 35 of *OpenAccess Series in Informatics (OASIcs)*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2013, pp. 89–94. URL: `http://drops.dagstuhl.de/opus/volltexte/2013/4276`. doi:`10.4230/OASIcs.ICCSW.2013.89`.

[44] D. Lopez-Paz, L. Bottou, B. Schölkopf, V. Vapnik, Unifying distillation and privileged information, arXiv preprint arXiv:1511.03643 (2015).

[45] S. Doncieux, D. Filliat, N. Díaz-Rodríguez, T. Hospedales, R. Duro, A. Coninx, D. M. Roijers, B. Girard, N. Perrin, O. Sigaud, Open-ended learning: A conceptual framework based on representational re-description, Frontiers in Neurorobotics 12 (2018) 59. URL: `https://www.frontiersin.org/article/10.3389/fnbot.2018.00059`. doi:`10.3389/fnbot.2018.00059`.

[46] S. Doncieux, Creativity: A driver for research on robotics in open environments, Intellectica 65 (2016) 205–219.

[47] J. M. Beer, L. Takayama, Mobile remote presence systems for older adults: acceptance, benefits, and concerns, in: Proceedings of the 6th international conference on Human-robot interaction, 2011, pp. 19–26.

[48] M. K. Ng, S. Primatesta, L. Giuliano, M. L. Lupetti, L. O. Russo, G. A. Farulla, M. Indaco, S. Rosa, C. Germak, B. Bona, A cloud robotics system for telepresence enabling mobility impaired people to enjoy the whole museum experience, in: 2015 10th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS), IEEE, 2015, pp. 1–6.

[49] F. Landi, L. Baraldi, M. Corsini, R. Cucchiara, Embodied vision-and-language navigation with dynamic convolutional filters, CoRR abs/1907.02985 (2019). URL: `http://arxiv.org/abs/1907.02985`. `arXiv:1907.02985`.

[50] V. Evers, N. Menezes, L. Merino, D. Gavrila, F. Nabais, M. Pantic, P. Alvito, D. Karreman, The development and real-world deployment of frog, the fun robotic outdoor guide, in: Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, 2014, pp. 100–100.

[51] V. Evers, N. Menezes, L. Merino, D. Gavrila, F. Nabais, M. Pantic, P. Alvito, The development and real-world application of frog, the fun robotic outdoor guide, in: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW Companion '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 281–284. URL: `https://doi.org/10.1145/2556420.2557638`. doi:`10.1145/2556420.2557638`.

[52] A. I. Karimov, E. E. Kopets, V. G. Rybin, S. V. Leonov, A. I. Voroshilova, D. N. Butusov, Advanced tone rendition technique for a painting robot, Robotics and Autonomous Systems 115 (2019) 17–27.

[53] J. F. J. Mellor, E. Park, Y. Ganin, I. Babuschkin, T. Kulkarni, D. Rosenbaum, A. Ballard, T. Weber, O. Vinyals, S. M. A. Eslami, Unsupervised doodling and painting with improved spiral, 2019. `arXiv:1910.01007`.

[54] F. Maillet, D. Eck, G. Desjardins, P. Lamere, et al., Steerable playlist generation by learning song similarity from radio station playlists., in: ISMIR, 2009, pp. 345–350.

[55] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A. M. Dai, M. D. Hoffman, D. Eck, An improved relative self-attention mechanism for transformer with application to music generation, ArXiv abs/1809.04281 (2018).

[56] A. Roberts, J. Engel, Y. Mann, J. Gillick, C. Kayacik, S. Nørly, M. Dinculescu, C. Radebaugh, C. Hawthorne, D. Eck, Magenta studio: Augmenting creativity with deep learning in ableton live, in: Proceedings of the International Workshop on Musical Metacreation (MUME), 2019. URL: http://musicalmetacreation.org/buddydrive/file/mume_2019_paper_2/.

[57] D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Human and automatic detection of generated text, arXiv preprint arXiv:1911.00650 (2019).

[58] M. Behrooz, J. Robertson, A. Jhala, Story quality as a matter of perception: Using word embeddings to estimate cognitive interest, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 15, 2019, pp. 3–9.

[59] M. Behrooz, A. Jhala, Modeling social interestingness in conversational stories, in: Proceedings of the Australasian Computer Science Week Multiconference, 2017, pp. 1–6.

[60] M. Behrooz, Curating Interest in Open Story Generation, Ph.D. thesis, UC Santa Cruz, 2019.

[61] K. Bollacker, N. Díaz-Rodríguez, X. Li, Extending Knowledge Graphs with Subjective Influence Networks for Personalized Fashion, Springer International Publishing, Cham, 2019, pp. 203–233. URL: https://doi.org/10.1007/978-3-030-00317-3_9. doi:10.1007/978-3-030-00317-3_9.

[62] T. Ramalho, T. Kociský, F. Besse, S. M. A. Eslami, G. Melis, F. Viola, P. Blunsom, K. M. Hermann, Encoding spatial relations from natural language, CoRR abs/1807.01670 (2018). URL: http://arxiv.org/abs/1807.01670. arXiv:1807.01670.

[63] M. Cornia, M. Stefanini, L. Baraldi, M. Corsini, R. Cucchiara, Explaining digital humanities by aligning images and textual descriptions, Pattern Recognition Letters 129 (2020) 166 – 172. URL: http://www.sciencedirect.com/science/article/pii/S0167865519303381. doi:https://doi.org/10.1016/j.patrec.2019.11.018.

[64] P.-Y. Oudeyer, Computational theories of curiosity-driven learning, arXiv preprint arXiv:1802.10546 (2018).

[65] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).

[66] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, N. Díaz-Rodríguez, Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges, Information Fusion 58 (2020) 52 – 68. URL: http://www.sciencedirect.com/science/article/pii/S1566253519307377. doi:https://doi.org/10.1016/j.inffus.2019.12.004.

[67] N. Pérez-Higueras, R. Ramón-Vigo, I. P. Hurtado, J. Capitán, F. Caballero, A social navigation system in telepresence robots for elderly, 2016.

[68] G.-Z. Yang, B. J. Nelson, R. R. Murphy, H. Choset, H. Christensen, S. H. Collins, P. Dario, K. Goldberg, K. Ikuta, N. Jacobstein, D. Kragic, R. H. Taylor, M. McNutt, Combating covid-19—the role of robotics in managing public health and infectious diseases, Science Robotics 5 (2020). URL: https://robotics.sciencemag.org/content/5/40/eabb5589. doi:10.1126/scirobotics.abb5589.