ARTICLE TYPE

XAITK: The Explainable AI Toolkit

Brian Hu | Paul Tunison | Bhavan Vasu | Nitesh Menon | Roddy Collins | Anthony Hoogs

Kitware, Inc., New York, USA

Correspondence
*Please address all correspondence to Brian Hu (Email: brian.hu@kitware.com) or Anthony Hoogs (Email: anthony.hoogs@kitware.com)

Summary

Recent advances in artificial intelligence (AI), driven mainly by deep neural networks, have yielded remarkable progress in fields such as computer vision, natural language processing, and reinforcement learning. Despite these successes, the inability to predict how AI systems will behave "in the wild" impacts almost all stages of planning and deployment, including research and development, verification and validation, and user trust and acceptance. The field of explainable artificial intelligence (XAI) seeks to develop techniques enabling AI algorithms to generate explanations of their results; generally these are human-interpretable representations or visualizations that are meant to "explain" how the system produced its outputs. We introduce the Explainable AI Toolkit (XAITK), a DARPA-sponsored effort that builds on results from the four-year DARPA XAI program. The XAITK has two goals: 1) to consolidate research results from DARPA XAI into a single publicly accessible repository, and 2) to identify operationally-relevant capabilities developed on DARPA XAI and assist in their transition to interested partners. We first describe the XAITK website and associated capabilities. These place the research results from DARPA XAI in the wider context of general research in the field of XAI, and include performer contributions of code, data, publications, and reports. We then describe the XAITK analytics and autonomy software frameworks. These are Python-based frameworks focused on particular XAI domains, and designed to provide a single integration endpoint for multiple algorithm implementations from across DARPA XAI. Each framework generalizes APIs for system-level data and control while providing a plugin interface for existing and future algorithm implementations. The XAITK project can be followed at: https://xaitk.org.

KEYWORDS:
explainable AI (XAI), toolkit, software framework, saliency, analytics, autonomy

1 | INTRODUCTION

Despite significant progress in the past few years, machine learning systems are still often viewed as "black boxes," which lack the ability to explain their output decisions to human users. In high-stakes situations such as healthcare, criminal justice, and autonomous driving [1,2,3], there is a need for explainable AI (XAI) tools that can help open up this black box. XAI is the field of machine learning that tries to make deep learning models more interpretable [4,5,6,7]. In doing so, XAI helps end users understand and appropriately trust machine learning-based systems, bridging AI capabilities and operational needs (Figure 1). This aligns

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ail2.40

well with the United States Department of Defense's push to adopt a set of five ethical principles for the development and deployment of autonomous systems[8]: responsible, equitable, traceable, reliable, and governable. The use of tools and resources from the field of XAI has the potential to inform many of these principles, and can help build appropriate user trust and understanding when designing and using autonomous systems.

Although several different taxonomies of explanation methods have been proposed, explanations generally fall into different categories based on their scope and mechanism. Local explanations provide interpretations of individual data points (e.g. images), while global explanations try to summarize models at the dataset level. Explanations can either be white-box or black-box, depending on the amount of access to the model the explanation requires. Black-box methods are model agnostic and can be applied more generally, while white-box methods often require the computation of model gradients. As an alternative to post-hoc explanation methods, models can also be made to be interpretable in the first place[2,3].

We propose a process for developing the Explainable AI Toolkit (XAITK). The goal of this toolkit is to capture and maintain the technical content developed under the DARPA XAI program[9,10], and to facilitate the transition of appropriate technology to interested partners. Current approaches, such as AI Explainability 360[11] (https://github.com/IBM/AIX360/) and InterpretML[12] (https://github.com/interpretml/interpret), focus on the researcher and deliver a unified experience of explainable AI capabilities via Jupyter notebooks. The XAITK project will complement this approach, engaging with transition partners to identify implementation pathways which directly address their needs. If successful, this project will provide a dynamic, evolving mechanism for continued application of explainable AI techniques to a wide array of problems and use cases across multiple domains.

In summary, the Explainable AI Toolkit (XAITK) has two main objectives: (1) to collect and curate the diverse research products of the four-year DARPA Explainable AI (XAI) program, and (2) to transition selected XAI capabilities to address the operational requirements of transition partners. The toolkit will contain XAI-specific capabilities and will also be a common, searchable framework that will provide scientific and technical guidance for understanding and deploying AI technology. Potential end users of the toolkit include transition partners within industry and government, researchers/engineers in the field of XAI, and other policy and decision makers. We believe that the XAI toolkit will be of broad interest to anyone who wants to deploy AI capabilities in operational settings and needs to validate, characterize and trust AI performance across a wide range of real-world conditions and application areas.

2 | OVERVIEW OF THE TOOLKIT

Next, we present an overview of the XAITK, including the proposed design process and toolkit components. In developing the XAITK, key technical challenges include the diversity of approaches and techniques for XAI, and providing technologies in a structure flexible enough to allow rapid development yet focused enough to integrate into production systems as easily as possible. We will address these challenges primarily via a structured, collaborative process allowing XAI team members to contribute artifacts and capabilities to an organized repository. We further plan to maximize the utility of the XAITK by developing software libraries providing XAI algorithms to clients via domain-specific APIs.

Framing the collective output of the DARPA XAI program as a toolkit gathers all program artifacts, including but not limited to software, into a single paradigm for inter-linked documentation, dissemination, and curation. As shown in Figure 2, non-software components such as publications, summary reports, and data sources can be viewed as "tools" which integrate at a different level from software, and whose user base is different from the typical software developer. However, these components were developed in the same ecosystem as XAI program software; folding them under the "toolkit" umbrella preserves that ecosystem. Some of these artifacts, such as publications and summary reports are essentially static. Others, such as data sources, may be managed by organizations outside the toolkit management purview. It is likely that the software and model components of the toolkit will be the most dynamic, and correspondingly require the most active management.

In developing the XAITK, we will work with the larger DARPA XAI community by creating an XAI Toolkit Working Group (XTWG). The XTWG will be responsible for the curation of different toolkit artifacts as well as the development of domain-specific software frameworks. Figure 3 diagrams the organizational structure of the working group in support of the overall XAI toolkit effort. The anticipated output of the toolkit and this working group includes: a set of XAITK Concept of Operations (ConOps) and Design Plan documents, the XAITK Artifact Repository (Sections 3.1 and 3.2), and software implementations of XAI technologies, suitable for integration into larger systems (Sections 3.5 and 3.6). The toolkit will also actively address transition efforts and strive to incorporate feedback from interested partners into the XAITK design process as early as possible (Section 4). We will describe each of these processes and components of the XAITK in more detail below.

3 | TOOLKIT COMPONENTS

3.1 | Non-software Artifacts

As discussed above, the toolkit will explicitly include non-software components. All of these artifacts will be stored in a comprehensive, centralized, user-accessible artifact repository that links related items. This repository will be accessible to users via a public-facing website: https://xaitk.org. We anticipate that subsequent management of these assets will involve periodic audits to ensure they are up-to-date. Assuming an artifact taxonomy similar to that in Figure 2, the Design Document will specify an appropriate "artifact lifecycle" for each class, addressing identification, acquisition, integration, documentation, and maintenance. For example, XAI publications would be reviewed for follow-on or related papers, datasets would be checked for accessibility or revisions, material related to subject studies would be updated as appropriate, and so on. Note that relative levels of effort required for each artifact class in Figure 2 are different. In particular, while datasets and publications are an integral part of the final XAITK, we anticipate that software will occupy most of the XTWG's attention.

3.2 | Software Artifacts

All software generated on the program should be, at a minimum, organized and checked into the toolkit. This would achieve the rightmost level of integration shown in Figure 4, in which the software exists as independent research modules. Figure 4 also shows possible further integration goals, ranging from developing a common API for functionally equivalent modules, through providing common system-level services such as GUI elements and data storage, all the way to complete integration within a unified prototype system. Choosing a target level of integration depends on many factors, but there is an inflection point where the balance tips from a "bottom-up" synthesis of XAI software into increasingly integrated modules, towards supporting a "top-down" goal of a complete system with XAI features.

3.3 | Benefits of Contributing to the Toolkit

As with most fields of machine learning, the field of explainable AI is somewhat fragmented, with papers and code scattered across multiple locations on the web. In contrast, the XAI toolkit offers a centralized and well-curated collection of papers, datasets, and code that also serve as artifacts tracking the history of the four-year DARPA XAI program. Figure 5 shows an overview of capabilities within the toolkit, which are searchable by either keyword tag or through an interactive concept map. Each capability contains associated paper and software resource pointers, as well as additional detailed metadata and information about the contribution, its intended use case, data and model considerations, as well as limitations. Both software and non-software contributions to the toolkit will help to further advance the field of explainable AI by collecting this research and knowledge into a central location and making it publicly available to a wide variety of users across academia, industry, and government. Additionally, by adding their capabilities to the toolkit, contributors: 1) can help support open-source, transparent, and reproducible XAI research, 2) benefit from searchability and discoverability of their contributions as part of a larger toolkit, and 3) go from research to production through possible transition to industry and government partners. Finally, by creating an open and relatively simple process for submitting contributions to the toolkit, we hope to ensure the long-term sustainability and adaptability of the toolkit as the field of XAI matures.

3.4 | Software Framework

Packaging implementations of research capabilities as toolkits facilitates potential partner evaluation and can highlight integration opportunities and challenges. Existing toolkits for XAI, such as AIX360[11], focus on enabling algorithmic exploration for the researcher, using Jupyter notebooks as an interface for analyses and visualization. This generic approach accommodates many XAI paradigms, but explicitly targets education and demonstration rather than transition or deployment. There is space for a complementary framework approach which provides systems and integration support for exploring a narrower set of algorithms that operate in more applied environments. Such a framework would act as middleware for selected classes of XAI algorithms,

providing not only common data types and APIs for more task-specific input and output modalities, but also supporting lifecycle operations required for application integration such as configuration, session management, and resource allocation. There is also opportunity for such a framework to assist in evaluation tasks, by providing a centralized location to collect and report various metrics. By creating a common software framework, specialized applications can natively access a suite of framework-enabled XAI implementations with minimal disruption to the application's original user interface and workflow.

We propose to develop an XAI framework pattern, which implements selected XAI techniques in domain-specific frameworks which abstract the algorithms and applications from each other. Each framework would be scoped to support a class of XAI techniques: for example, one framework might support saliency maps via perturbation algorithms; another might support explainable reinforcement learning. This framework pattern is divided into two layers, the general XAI Domain Descriptor and the specific Domain Implementations. The Domain Descriptor is a mixture of human- and application-interpretable metadata describing specific domain implementations, such as free-form text describing the domain, data types, visualization and interaction methods, and so forth. These capabilities are provided by the software in the Domain Implementation. Separating the overall XAI algorithm into a generalized descriptor and a domain-specific implementation allows a measure of discoverability through methods such as automatically-generated documentation. This approach also allows for multiple levels of abstraction and visibility for any given XAI capability. For example, a researcher or developer may require different information about a capability compared to someone who is responsible for planning, assessment, or evaluation. The presentation of information at multiple levels allows end users to evaluate whether or not to use an algorithm, focusing on implementation details only when necessary.

3.5 | Analytics Domain Implementation

Saliency maps are a form of visual explanation which attempts to provide the user insight into which input regions the model pays attention to during the prediction process. For computer vision models, various saliency algorithms have been proposed as forms of XAI which can highlight regions of the input image responsible for the model's output decision [13,14,15,16]. While most XAI techniques in this area have been developed for the task of image classification (e.g. Grad-CAM [15]), there has been an increasing push to create explanations for other image understanding tasks, including object detection [17] and image similarity [18,19,20,21,22]. A notional architecture diagram for an analytics domain implementation is shown in Figure 6, here supporting saliency maps as an example. On the client side (right), the framework provides hooks for configuration, supplying lists of images, receiving lists of saliency maps, etc. On the XAI algorithm side (left), the framework provides hooks for startup and shutdown of individual algorithm instances, access to raw pixel data, etc. The framework insulates the client from allocating and managing algorithm instances, and insulates the algorithm developer from issues such as configuration, accessing and decoding client data, and disposition of algorithm output.

3.6 | Autonomy Domain Implementation

Deep reinforcement learning seeks to learn generalizable action policies that can control autonomous systems in different environments. Several XAI techniques have been developed to explain the behavior of these deep reinforcement learning systems [23,24,25,26]. Figure 7 shows a notional diagram of an example autonomy domain implementation, enabling exploration of critical state algorithms [23] for reinforcement learning as an example. Here we assume the client application is specialized to the task at hand (e.g. a video game, self-driving vehicle, etc.) and that the framework is enabling a comparison between different algorithms.

4 | TOOLKIT TRANSITION EFFORT

In parallel with toolkit development, we anticipate that transition efforts will be identifying and reaching out to potential partners. As described below, we expect these interactions to influence XAITK design and implementation at several levels as work progresses. By working closely with interested partners, the XAITK will be able to identify domains, datasets, and workflows which could benefit from XAI and the XAITK. Toolkit development will be driven by these use cases, ensuring that the research code generated by the DARPA XAI program can be productionized and used in deployed systems. Figure 8 shows potential use cases of XAI with varying points of system integration. Explanations can be presented for the underlying data as a form of preprocessing, help create interpretable models as part of the system itself, or be presented to users in a post-hoc manner, such as

for methods like saliency maps. Combinations of each of these approaches can also be used. Identifying the level at which XAI is used is critical for understanding potential use cases and the anticipated level of system integration required.

We anticipate that the transition effort will include outreach and education, development, dataset integration, codebase integration, and evaluation on partner data. Transition of the XAITK will require identifying likely partners, and engaging with them to describe current and near-term XAI capabilities. We will also explore opportunities where XAI could benefit these partners. Following identification of interested partners, detailed analysis of partner workflows to identify specific scenarios where XAI can help and estimation of the work required vs. expected benefit, resulting in a decision whether or not to proceed further. If there is continued interest, the XAITK will work closely with the partner's existing data and software infrastructure. The partner provides relevant data, which could be used to assess XAITK performance, train models, and highlight software engineering issues required for further integration (such as file formats, metadata, etc.) There will be continued in-depth discussions of software integration pathways and requirements. The XAITK team will have to understand details on software environments, especially for working on non-public or restricted data. XAITK will also identify possible avenues for demonstrations and evaluations to allow broad dissemination of the capabilities present within the toolkit (regardless of integration level). As a final goal, we will also pursue formalized evaluation of XAITK capabilities on partner-supplied data, either qualitative (with ground-truth) or quantitative (using SME / analyst assessment, questionnaires, etc.) In all of these efforts, the XTWG will serve as an active liason between transition partners and DARPA.

5 | DISCUSSION

We believe the XAI toolkit will be a valuable asset to the larger explainable AI community within academia, industry, and government. We envision a wide array of users, ranging from researchers, model developers, system integrators, to policymakers and other stakeholders. The toolkit fills a critical need when transitioning and deploying AI systems by providing an open-source collection of state-of-the-art explainable AI tools and resources, initially building off prior work from the DARPA XAI program. Importantly, the toolkit aims to cover the entire AI lifecycle, with a set of techniques that can integrate with data and models at the pre-processing, design, training, and operation stages, all the way through to the final testing and evaluation of models. Due to its open-source nature, the toolkit is easily extendable to new algorithms and findings as the field of XAI evolves, which will also help with its long-term sustainability, growth, and community adoption.

Building a sustainable, open-source XAI toolkit is a huge undertaking. Risks include lack of participation from DARPA XAI performers in the XAITK design process; failure by those performers to integrate contributions into a coherent repository; and potential transition partners not engaging with the program. We will address the first two risks via early outreach to the DARPA XAI community and making it clear that we view XAITK development as a collaborative, transparent process. We will leverage the work already done by the program office soliciting potential contributions from the community. The risk of lack of engagement by transition partners will also be addressed by early outreach, assisted by other contacts through the program office.

The XAITK effort represents an ambitious effort to not only collect and curate the various software and non-software artifacts from the DARPA XAI program, but to also provide domain-specific software frameworks that integrate multiple analytics and autonomy algorithms. The hope is that these algorithms can then be transitioned to different partners, helping bridge nascent AI capabilities and real-world operational needs. Finally, the XAITK effort represents a larger process and is not simply a product. We envision the XAITK to be an open-source resource that will benefit the larger XAI community by collecting research and lessons learned from the field of XAI. We welcome the XAI community's participation and feedback in this effort.

ACKNOWLEDGMENTS

This material is based on research sponsored by Air Force Research Laboratory and DARPA under Cooperative Agreement number N66001-17-2-4028. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory and DARPA or the U.S. Government. Distribution Statement 'A' (Approved for Public Release, Distribution Unlimited).

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

References

1. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?. *arXiv preprint arXiv:1712.09923* 2017.
2. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* 2017.
3. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 2019; 1(5): 206–215.
4. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* 2017.
5. Adadi A, Berrada M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 2018; 6: 52138–52160.
6. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 2020; 58: 82–115.
7. Vilone G, Longo L. Explainable Artificial Intelligence: a Systematic Review. *arXiv preprint arXiv:2006.00093* 2020.
8. Board DI. AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. *Supporting document, Defense Innovation Board* 2019.
9. Gunning D. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2017; 2(2).
10. Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine* 2019; 40(2): 44–58.
11. Arya V, Bellamy RKE, Chen PY, et al. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. 2019.
12. Nori H, Jenkins S, Koch P, Caruana R. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* 2019.
13. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Springer. ; 2014: 818–833.
14. Ribeiro MT, Singh S, Guestrin C. " Why should I trust you?" Explaining the predictions of any classifier. In: ; 2016: 1135–1144.
15. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ; 2017: 618–626.
16. Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: ; 2017: 3429–3437.
17. Petsiuk V, Jain R, Manjunatha V, et al. Black-box Explanation of Object Detectors via Saliency Maps. *arXiv preprint arXiv:2006.03204* 2020.
18. Dong B, Collins R, Hoogs A. Explainability for Content-Based Image Retrieval.. In: ; 2019: 95–98.

19. Stylianou A, Souvenir R, Pless R. Visualizing deep similarity networks. In: IEEE. ; 2019: 2029–2037.

20. Williford JR, May BB, Byrne J. Explainable Face Recognition. In: Springer. ; 2020: 248–263.

21. Eberle O, Büttner J, Kräutli F, Müller KR, Valleriani M, Montavon G. Building and Interpreting Deep Similarity Models. *arXiv preprint arXiv:2003.05431* 2020.

22. Chen L, Chen J, Hajimirsadeghi H, Mori G. Adapting Grad-CAM for embedding networks. In: ; 2020: 2794–2803.

23. Huang SH, Bhatia K, Abbeel P, Dragan AD. Establishing appropriate trust via critical states. In: IEEE. ; 2018: 3929–3936.

24. Iyer R, Li Y, Li H, Lewis M, Sundar R, Sycara K. Transparency and explanation in deep reinforcement learning neural networks. In: ; 2018: 144–150.

25. Noothigattu R, Bouneffouf D, Mattei N, et al. Interpretable multi-objective reinforcement learning through policy orchestration. *arXiv preprint arXiv:1809.08343* 2018.

26. Atrey A, Clary K, Jensen D. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. *arXiv preprint arXiv:1912.05743* 2019.

List of Figures

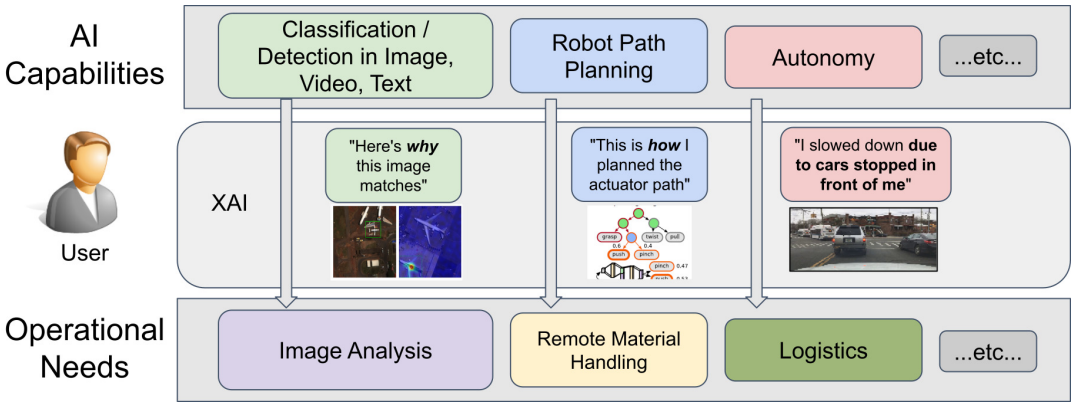


FIGURE 1 How XAI can help AI help users. XAI serves as a critical bridge between state-of-the-art AI capabilities and existing operational needs, helping to ensure appropriate user trust and acceptance of models. XAI can be applied to different use cases including image analysis, remote material handling, and logistics, which can involve both analytics and autonomy applications.

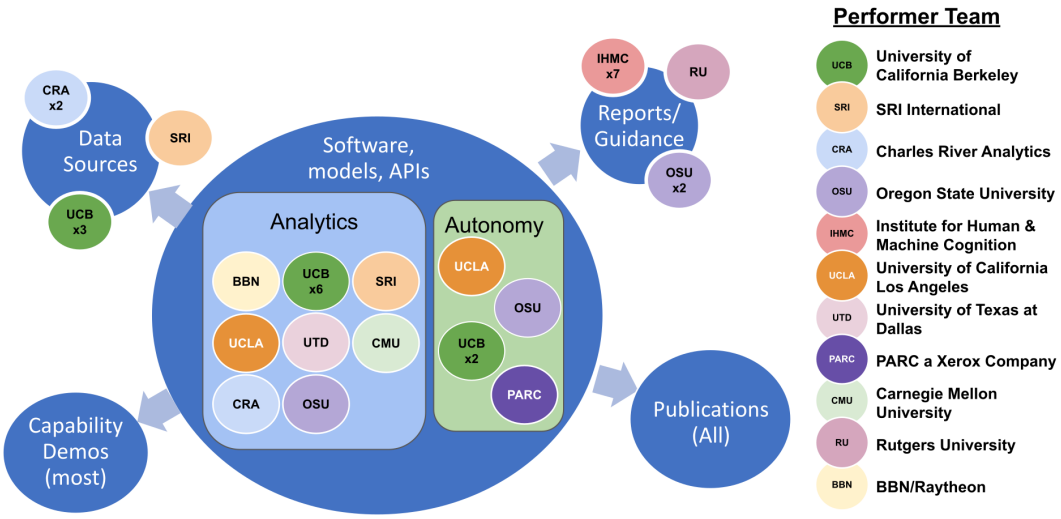


FIGURE 2 Components of the XAI Toolkit address a spectrum of program artifacts. Non-software artifacts include publications, reports and guidance, and data sources. Software artifacts include domain-specific software frameworks (e.g. for analytics and autonomy) as well as demos of different XAI capabilities. The diagram represents the anticipated toolkit contributions from the different DARPA XAI performer teams shown to the right.

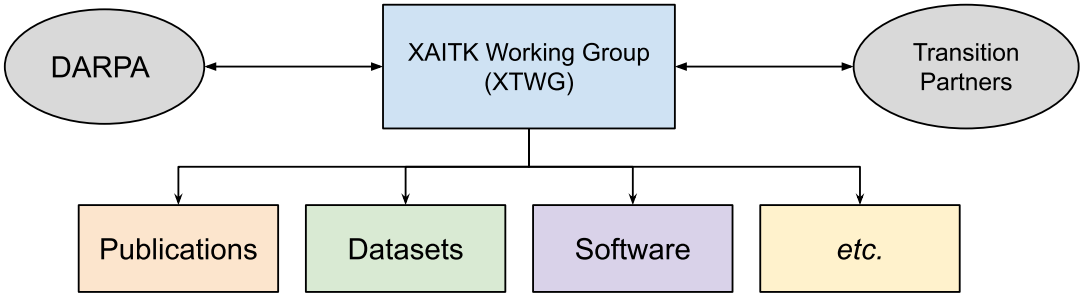


FIGURE 3 Organizational structure for the XAI Toolkit Working Group (XTWG). The XTWG (blue box) is notionally responsible for collecting and curating the different artifacts (e.g. publications, datasets, software, etc.) from the DARPA XAI program (bottom row). The XTWG also acts as a liason between DARPA and potential transition partners (gray ovals), identifying relevant XAI use cases that can inform toolkit design.

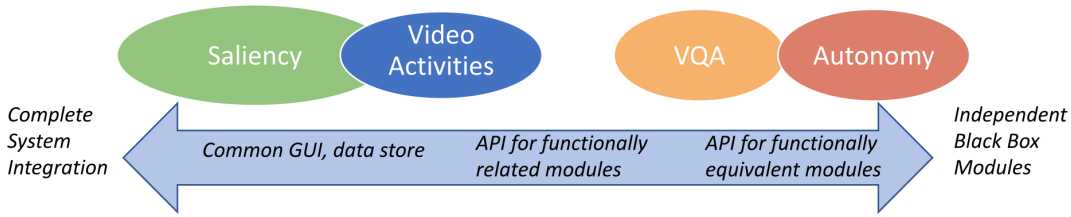
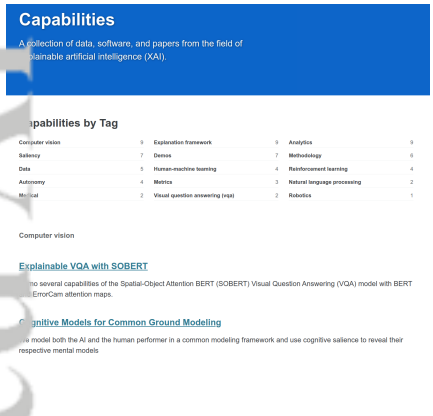
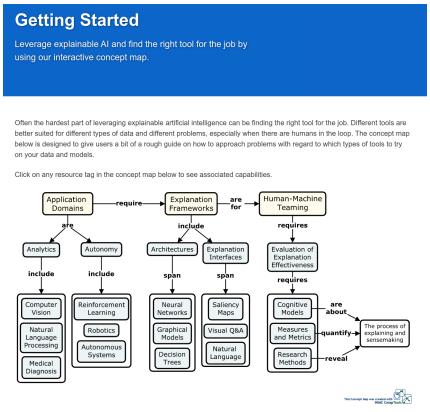


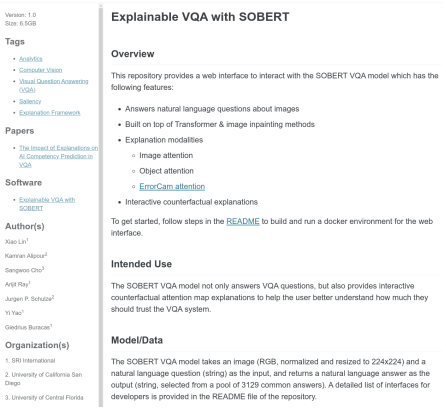
FIGURE 4 Degrees of software integration. Planned contributions to the XAI toolkit span a wide spectrum, ranging from independent modules to completely integrated systems. Based on an initial assessment of these contributions, we have grouped together certain contributions and placed them along the software integration spectrum. For example, there are many planned contributions around saliency, making a common software framework possible.



(A) Capabilities by tag



(B) Capabilities by concept map



(C) Example capability

FIGURE 5 Overview of capabilities within the XAI toolkit (viewable on https://xaitk.org). (A) Capabilities are searchable and grouped together by keyword tag. (B) Capabilities are also organized using an interactive concept map, linking related sets of capabilities and guiding users on how to select the appropriate tool or resource. (C) An example capability is shown with paper and software resource pointers to the left and associated metadata.

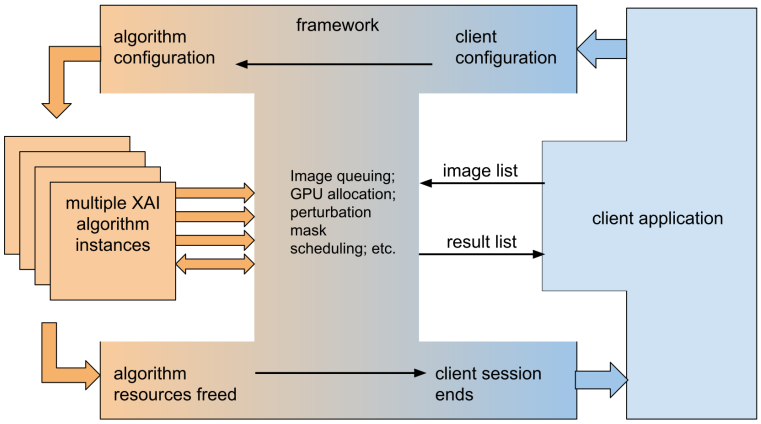


FIGURE 6 Notional architectural framework for saliency maps based on image perturbations. On the right, the client supplies configuration information (including algorithm selection and resource identifiers such as GPU limits, image access information, etc.) On the left, the XAI algorithm developer has coded against an API supplying raw image data. In the middle, the framework handles requests for multiple results by launching per-image instances of the XAI algorithms, ensuring GPU limits are respected, providing status information to the client.

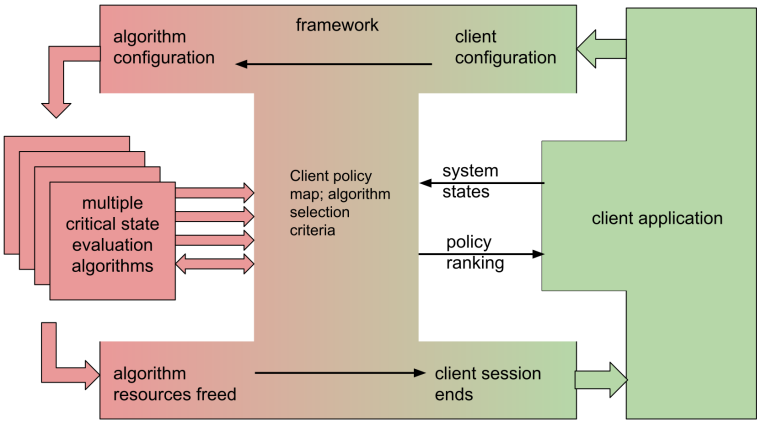


FIGURE 7 Notional architectural framework for critical state evaluation[23] of deep reinforcement learning systems. The implementation follows a similar pattern as that for the analytics domain, consisting of a client application which handles configuration and API requests that allows for the evaluation of multiple XAI algorithms.

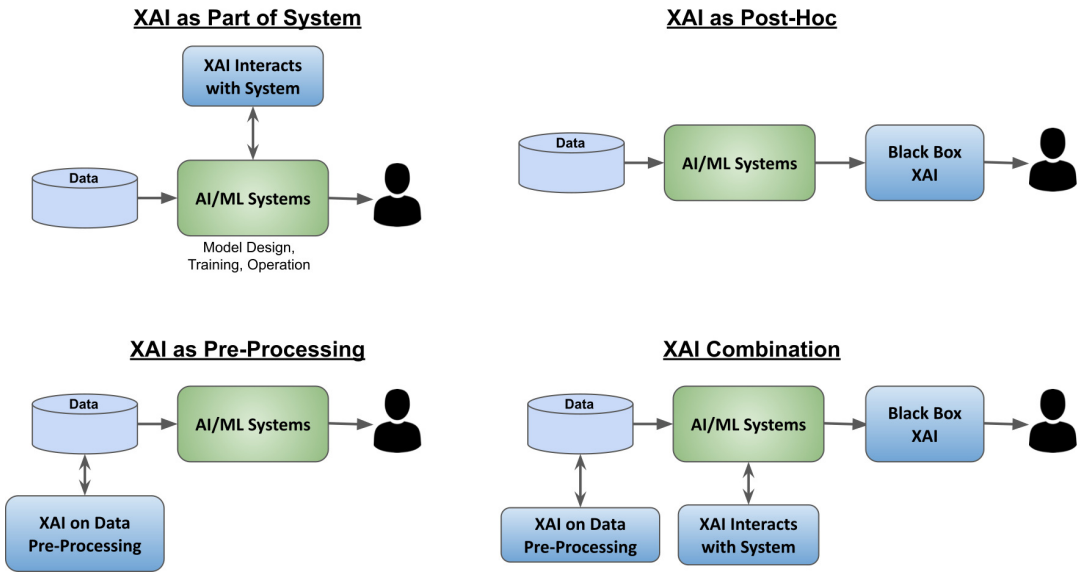


FIGURE 8 XAI system integration points. Each diagram shows a different XAI use case (clockwise, starting from the lower left): XAI as pre-processing (e.g. explaining the data), XAI as part of the system (e.g. inherently interpretable models), XAI as post-hoc explanation (e.g. visual saliency maps), and XAI as a combination of all the previous methods.