A Modular Interface Framework for Multimodal Annotations and Visualizations in Human-AI Collaboration

by

Chris Kim

A thesis submitted to the
School of Graduate and Postdoctoral Studies in partial
fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

Faculty of Science
University of Ontario Institute of Technology
(Ontario Tech University)
Oshawa, Ontario, Canada
July 2021

THESIS EXAMINATION INFORMATION

Submitted by: **Chris Kim**

Doctor of Philosophy in **Computer Science**

Thesis title: A Modular Interface Framework for Multimodal Annotations and Visualizations in Human-AI Collaboration

An oral defense of this thesis took place on April 8, 2021 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Dr. Stephen Marsh
Research Supervisor	Dr. Christopher Collins
Examining Committee Member	Dr. Jaroslaw Szlichta
Examining Committee Member	Dr. Mohamed Amer
University Examiner	Dr. Ken Pu
External Examiner	Dr. Enrico Bertini, NYU

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Abstract

Modular is a web-based annotation, visualization, and inference software platform for computational language and vision research. The platform enables researchers to set up an interface for efficiently annotating language and vision datasets, visualizing the predictions made by a machine learning model, and interacting with an intelligent system.

Artificial intelligence (AI) research, including machine learning, computer vision, and natural language processing, requires large amounts of annotated data. The current research and development pipeline involves each group collecting their own datasets using an annotation tool tailored specifically to their needs, followed by a series of engineering efforts in loading other external datasets and developing their own interfaces, often mimicking some components of existing annotation tools. Extensible and customizable as required by individual projects, the framework has been successfully applied to a number of research efforts in human-AI collaboration, including commonsense grounding of language and vision data, conversational AI for collaboration with human users, and explainable AI in improving interpretability of the AI system.

Facilitated by the aforementioned Modular framework, the dissertation examines a notable set of opportunities that inspire the new, productive symbiosis between human users and AI agents, where the two parties can successfully complete a complex task together and mutually benefit in providing advantages missing from the other party. Finally, the dissertation sets out to evaluate whether human users can establish a level of appropriate trust and reliance through AI explanation.

Keyword: artificial intelligence; interactive visualization; machine learning; knowledge representation; human-computer interaction

Author's Declaration

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Ontario Institute of Technology (Ontario Tech University) to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Ontario Institute of Technology (Ontario Tech University) to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

The research work in this thesis was performed in compliance with the regulations of Research Ethics Board under File Number 14975.

_____ Chris Kim

Statement of Contributions

I hereby declare that I am the sole author of this dissertation, and that I am the sole source of the creative works and/or inventive knowledge described in this document. Ideas and figures described herein have been previously made public in the following venues:

[CV1] M. R. Amer, T. J. Meo, A. N. Raghavan, A. C. Tozzo, A. Tamrakar, D. A. Salter, and K.-Y. Kim. Artificial intelligence in interactive storytelling, Sept. 29 2020. US Patent 10,789,755.

[CV2] C. Kim. A Mixed Initiative Platform for: Explanation, Conversation, and Commonsense in Movies. In *IRAD Poster Day Social, SRI International*, 2018.

[CV3] C. Kim. A modular framework for collaborative multimodal annotation and visualization. In *International Conference on Intelligent User Interfaces: Companion*, pages 165–166, 2019.

[CV4] C. Kim, X. Lin, C. Collins, G. W. Taylor, and M. R. Amer. Learn, Generate, Rank, Explain: A Case Study of Visual Explanation by Generative Machine Learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2021.

[CV5] X. Lin, C. Kim, T. Meo, and M. R. Amer. Learn, Generate, Rank: Generative Ranking of Motion Capture. In *European Conference on Computer Vision*, 2018.

[CV6] T. J. Meo, C. Kim, A. Raghavan, A. Tozzo, D. A. Salter, A. Tamrakar, and M. R. Amer. Aesop: A visual storytelling platform for conversational AI and common sense grounding. *AI Communications*, 32(1):59–76, 2019.

[CV7] D. Quick, D. Burke, and C. Kim. MUSICA: Musical Interactive Collaborative Agent. In *DARPA CwC PI Meeting*, 2020.

[CV8] D. Quick, D. Burke, and C. Kim. MUSICA: Musical Interactive Collaborative Agent. In *Conference on AI Music Creativity, AIMC*, 2021.

Acknowledgements

I wish to express my gratitude to my examining committee members — Dr. Christopher Collins, Dr. Jaroslaw Szlichta, Dr. Mohamed Amer, Dr. Enrico Bertini, Dr. Ken Pu, and Dr. Stephen Marsh — for their unwavering support over the years. I also wish to thank the Faculty of Science at the Ontario Tech University for fostering an ideal environment for our academic endeavour.

This dissertation, along with my academic journey, simply would not have been possible without my academic supervisor Dr. Christopher Collins, who graciously granted me an opportunity to join his research lab and provided the necessary support and guidance every step of the way. Dr. Collins is a reliable guide, an attentive teacher, and an ever-inspiring mentor — and I look forward to opportunities to collaborate once again with Dr. Collins in the future.

Throughout this primarily solitary experience, I found solace in conversations and coffee breaks with my family, colleagues and friends. Thank you for being a part of my life, in one way or another.

Table of Contents

List of Tables

Tables

List of Figures

Figures

Chapter 1

Introduction

I saw – with shut eyes, but acute mental vision – I saw the pale student of unhallowed arts kneeling beside the thing he had put together. I saw the hideous phantasm of a man stretched out, and then, on the working of some powerful engine, show signs of life and stir with an uneasy, half-vital motion. Frightful must it be, for supremely frightful would be the effect of any human endeavour to mock the stupendous mechanism of the Creator of the world.

Mary Shelley, *Frankenstein; or, The Modern Prometheus* [144]

1.1 Motivation

The story of artificial beings imitating human characteristics remains a strong object of fascination since classical antiquity, inspiring products that spark policy and ethical debates of today. As the legendary sculptor Pygmalion carves his spouse out of ivory in Greek myths, the Spike Jonze film *Her* tells a story of a lonely man who falls in love with his voice-activated personal assistant. Battle of wits ensues between man and machine as the Mechanical Turk tours around the 19th century Europe, only to turn out to be a real player-in-the-box; meanwhile, the world watches as IBM's Deep Blue, an actual supercomputer, wins against the chess grandmaster Garry Kasparov at the turn of the millennium. Finally, as Victor Frankenstein desperately runs away his own monstrous, nameless creation in Mary Shelley's eponymous novel, we eagerly ask Amazon Alexa

and other smart speakers about today's weather and news as we pour ourselves some morning coffee. Beyond these fascinating stories of love, fear, and competition, however, we now live in the new reality of consistently interacting with AI products on a daily basis.

With AI assistance firmly ingrained in our professional and social activities, ranging from composing friendly email messages with intelligent word suggestions to reviewing judicial precedents at an unprecedented scale, it is difficult to imagine human-computer interaction without such tools in our daily lives; onlookers dismiss these tools as "mainstream" as they turn their attention to more exciting, cutting edge applications, with their minds reeling in excitement over autonomous vehicles and personal assistants with more human-like qualities. In any case, artificial intelligence to its fans is a promise of efficiency and scalability, as well as an object of consistent reliance.

On the other hand, critics of artificial intelligence warn of a dystopian future, fueled by our submission to AI and surveillance states that assume control over such technologies. Beyond the infamous yet seemingly old-fashioned network of traditional closed circuit cameras, the contemporary Chinese government proudly presents its system of AI algorithms that censor communication and coerce its own citizens into submission [89]; the recent controversy of an American facial recognition solution enterprise Clearview AI, amassing its dataset based on publicly available social network data and then licensing its solution to prominent law enforcement across North America [137], paints a renewed picture of Orwell's vision state surveillance and "Big Brother." With self-driving car fatalities and automation-induced labour market anxieties entering today's conversations, the AI critics conclude that artificial intelligence should be subject to strong scrutiny and regulation, let alone fear and suspicion.

Despite the seemingly ever-present discourse surrounding human users and artificial intelligence, as well as the recent physical proximity between the two, the relationship continues to be an anxious, uncomfortable one. While users are happy to take advantage of capable powerful AI solutions, they also recognize their political implications; while researchers continue to improve AI algorithms with additional datasets and calibrations, they consider the ethical need to ade-

quately constrain and control them. It is abundantly clear that humans and AI rely on each other, where AI solutions depend on human data and supervision, and humans use such solutions for productivity. Why must their relationship be a contradictory one, built on fascination and fear of displacement?

Over the past few decades, there seems to have been a series of efforts to extend beyond mere coexistence between the two parties and establish a collaborative relationship. As the users become acclimatized with AI applications ranging from anthropomorphic chatbots to complex forecasting networks, demands for more human-like behaviour and user-friendly features emerge. Users may request their AI partners to be less operationally complex and more convincing in their outputs, and various stakeholders in the human-computer interaction (HCI) community set out to address such challenges from various angles, including improved accuracy and speed, ergonomic interface design, even the system's attempts to explain itself algorithmically. These efforts just may bridge the trust chasm once and for all, or further dampen the excitement.

1.2 Dissertation Overview

Beyond the present polarizing sentiments, this dissertation sets out to examine a notable set of opportunities that inspire the new, productive interdependence of human users and AI agents. In this complementary relationship, the two parties can successfully complete a complex task together, mutually benefit in providing advantages missing from the other party, and the human users can establish a level of trust and reliance through AI explanation. Finally, this dissertation presents an adaptable interface framework designed for practitioners or researchers who wish to quickly generate and deploy projects that capture such opportunities — complete with its technical blueprint and use cases.

1.2.1 Challenges in Human-AI Collaboration

Since the introduction of consumer-level AI technologies and business applications featuring AI assistance, such implementations have been the consistent subject of fascination and scrutiny. From Microsoft Office's Clippy to Amazon

Alexa, many consumer products seek to imitate human interactions with some success, while businesses are eager to incorporate AI insights as part of their decision making processes. Despite this enthusiasm, it is difficult to consider AI a genuine partner rather than a mere tool designed to facilitate individual tasks: AI assistants are yet to understand the "bigger picture" of the user's or the organization's overarching objective and respond proactively as a teammate.

Chapter 2 presents a brief overview of past and present challenges surrounding development and deployment of AI systems with collaborative capabilities. Featuring a range of research areas including application design, task development, and institutional adoption, the chapter defines the current gaps in facilitating human-AI collaboration and sets the stage for the subsequent chapters in this dissertation.

1.2.2 Multimodal Annotation and Visualization Platform

As the demand for more accessible, extensible AI systems emerge, there is an opportunity to provide a more flexible solution that accommodates a variety of needs and stakeholders: a research group may want to collect user-generated data from a select group of annotators, or demonstrate a novel AI model to a larger audience with a series of visualizations; the same group many also want to deploy and conduct a large user-study to evaluate the same AI model. While there are many out-of-the-box solutions that respond to such needs, many research groups resort to building new tools specifically designed for their project objectives — a costly endeavour prone to failures and defects common in software development.

Chapter 3 presents a novel software platform to establishing a common framework that enables an end-to-end experience of annotating datasets, visualizing AI output, and deploying user studies, culminating in an open-source interface platform that accommodates different types of data annotations and supports relationships between such annotations using a graph data structure. Fueled by its modular and extensible mechanism where individual users can combine different modules for different project needs, this platform serves as a core foundation to work presented in the subsequent chapters.

1.2.3 Annotation for Commonsense Reasoning

Committed to simulating the human ability to recognize ordinary situations and react accordingly through the process of extrapolation and judgment, the commonsense grounding branch of research has a seemingly simple goal of training AI in the fundamentals of basic human understanding of the physical world. Ideal systems built on such insights would be able to make inferences and decisions that imitate humans in similar situations, and naturally, building such systems remains a remarkable challenge: a large-scale acquisition of knowledge and behaviour from humans, in a form of text-based queries, image descriptions, and movements in a digital or physical space.

In Chapter 4, the dissertation presents *Aesop*: a visual storytelling platform that demonstrates the ability to collect commonsense knowledge as derived from the cinematic history and represent them as novel knowledge graphs. Using a full-length film as the starting point of the annotation process, Aesop allows its users to watch and represent the entire film as a collection of on-screen entities, text labels, bounding boxes, and more — all tightly linked as a spatio-temporal graph representation of individual scenes.

1.2.4 Collaboration for Complex Tasks

While some worry about displacement of human workers due to the emergence of AI-driven automation, other critics claim that some areas — such as creative industries — are immune to such dangers [61]. Despite this sense of rivalry, there exist complex tasks that both parties must be involved to complete with speed and ease: a time-consuming task of investigating an overwhelming number of court case documents could benefit from human-AI collaboration, as the human user manually reviews the shortlist of documents identified by AI; a creative endeavour of designing a new grocery store can be facilitated by using an AI agent that the interior designer can speak to.

The dissertation addresses such challenges and presents two case studies in favour of human-AI collaboration. In Chapter 5, Aesop once again presents an opportunity for the user to take the helm of a film director: based on previously established knowledge graph and commonsense reasoning, the user can interact

with an AI-based chatbot to build a scene, place actors, and ask for performance. In addition, the chapter explores another creative task of generating music with the help of AI assistance and improvising in real-time, trading jazz solos.

1.2.5 AI Explanation for Human Understanding

As alluring and inviting as the promise of artificial intelligence may be, there is an inherent tension between performance and explainability in today's popular AI systems. While built on human intelligence to assist with human needs, various components remain opaque in their current behaviour and future outcome: when the user receives a specific decision from the conventional "black-box" AI system, it is difficult to explain how the system arrived at this answer, and extrapolate how the system will continue to perform in the future. Seeking to bridge this gap are Explainable AI (XAI) systems that aim to maintain performance of AI systems while providing rationale for its decisions. A successful XAI system should empower users to understand its strengths and weaknesses and even correct the system's mistakes to improve its future performance.

Inspired by these opportunities, Chapter 6 introduces the paradigm of *explanation by generation*, where the novel XAI system tackles the prominent challenge of visual search and ranking using a generative framework based on the Dense Validation Generative Adversarial Networks (DVGANs) approach. In a typical visual search and ranking challenge, the system is given a query and the goal to retrieve videos that contain the query. The XAI system approaches this problem by generating multiple visual hypotheses based on the query and searching the videos using these evidences. By providing the user a clear interactive visualization interface that supports the usual benefits of the AI system and provides the model-generated videos, the XAI system essentially explains its decision-making process by showing what the system "thinks" the answer should be.

1.2.6 Trust and Reliance Between Human and AI

Despite the belief that a more explainable, less opaque XAI system will garner a higher level of trust and reliance between human users and AI, it is unclear

whether the XAI system will truly facilitate the user's understanding of the AI model and improve task performance. The user may find the XAI system's explanations disruptive, no matter how "benevolent" one may perceive such assistance. Instead, the user may prefer the black-box AI system for its opaque yet speedier and more accurate responses [65].

Chapter 7 sets out to evaluate how the generative XAI system fares in comparison to the conventional AI system using three main criteria: user's mental model of the AI system, task performance alongside the system, and appropriate trust and reliance exemplified by the user's confidence in the system. Prompting the users to tackle three distinct categories of challenges that involve a large set of video clips, the user study sets out to assess the benefits of the XAI system.

1.3 Summary

This text presents the following chapters in support of human-AI collaboration:

- **Chapter 2:** Overview of present challenges in human-AI collaboration
- **Chapter 3:** Modular interface framework for collecting multimodal annotations, visualizing AI models, and deploying large-scale user studies
- **Chapter 4:** Application of the framework to collecting annotations for commonsense grounding
- **Chapter 5:** Case for human-AI collaboration in completing complex mixed-initiative tasks
- **Chapter 6:** Approach to building more explainable AI models using generative ranking
- **Chapter 7:** Experiment for assessing user's trust and reliance in human-AI collaboration — and the findings

Finally, **Chapter 8** summarizes the previous chapters and proposes opportunities and directions for future research and application enabled by contributions outlined in the dissertation.

Chapter 2

Challenges in Human-AI Collaboration

Our coexistence with artificial intelligence hinges on combining what is humanly unattainable — the hugely scaled narrow AI intelligence that will only get better at any given domain with what we humans can uniquely offer to one another ... Narrow AI has no self awareness, emotions, or a heart. Narrow AI has no sense of beauty, fun, or humor. It doesn't even have feelings or self-consciousness. Can you imagine the ecstasy that comes from beating a world champion? AlphaGo bested the globe's best player, but took no pleasure in the game, felt no happiness from winning, and had no desire to hug a loved one after its victory.

Kai-Fu Lee, *A Blueprint for Coexistence with Artificial Intelligence* [85]

2.1 Introduction

Recent progress in AI research presents an exciting vision of the world where human users and machine agents can actively communicate and collaborate with each other [96], yet today's available solutions offer little more than chatbots and smart speakers passively standing by for user requests. In order to support collaboration as a fellow teammate, AI must interact with their human partners in meaningful, significant ways by being able to engage in the process of

solving a complex problem: it must be able to define a problem area, propose potential solutions, and learn from past interactions beyond mere automation of actions [143].

Meanwhile, the success of such collaborative efforts hinges on factors beyond capabilities of AI and AI-activated applications that users interact with. Poorly designed tasks, inadequate support systems, and unclear objectives detract effective teamwork even amongst human collaborators [26], and one must design work practices conducive to involving multiple parties even before bringing in a hypothetical AI partner.

Finally, it is unclear whether such proactive, sophisticated AI assistants will be embraced by their human partners and organizations they serve. AI may add tremendous value to existing teams in augmenting their performance and benefits [163], even participating in co-creation alongside human users in mixed-initiative tasks [165]; meanwhile, its presence also may be seen as a threat to the safety of human partners and the society at large. Some pose a more optimistic vision of AI taking over some of the more unpleasant tasks, yet there also seems to be a necessity for large organizational change and deeper understanding of AI training for successful business adoption [131]. In addition, whether a human user will appropriately trust and rely on such intelligent machine partners remains to be seen [84], especially in mixed-initiative tasks where human and machine agents take turns in completing an objective together [90].

Inspired by emergent design areas in human-AI collaboration [143], this chapter briefly introduces various challenges found in applications that support such joint activities, tasks that are conducive to collaboration, and efforts in formalizing procedures of AI adoption. In following the trajectory of these various developments, this chapter sets out to establish emerging opportunities that the dissertation explores in subsequent chapters.

2.2 Limitations of Existing Systems

Primarily designed to facilitate teamwork between human stakeholders, many design features and organizing frameworks featured in today's collaborative ap-

plications fall short when applied to support human-AI interaction. The collaborative process is inherently "incremental, subject to negotiation, and forever tentative," and the ideal collaborative AI system must support interaction that embodies the "kind of give-and-take" common in natural teamwork among groups of people [13]: each visual and behavioral attribute found in the AI system must be scrutinized and calibrated for successful teamwork experience.

The importance of the machine teammate's appearance is often discounted and considered secondary to objective performance, and its visual design and perceived personality are generally left to the realms of science fiction and popular culture. However, a machine teammate's appearance does indeed play a significant role in the human participant's perception of "likeability, engagement, trust and satisfaction" [86] and affects one's willingness to comply with its instructions and cooperate [52]. As AI solutions become more increasingly integrated in the social realm, they are now closer to being considered friendly companions that reflect "common assumptions and views" of the era [27]. Beyond the initial novelty of interacting with such artefacts, their synthetic agents are much more prone to criticism in terms of their visual fidelity, affective capabilities, and the very social values they embody [10] — presenting an opportunity for systems that can flexibly adjust their presentation in response to changing needs of the users.

Despite concerns surrounding of data privacy and residency, the emerging domain of knowledge processing and machine learning has become a staple to a generation of AI systems powered by human insights and body of knowledge, digitized and stored for automation and replication. Though, a number of challenges emerge upon applying the lens of human-AI collaboration to this domain. With human users, the AI system ideally should be able to differentiate between serious requests and idle actions exhibited by their teammates, and adapt its behavior based on new findings. While conventional consumer solutions such as Google Nest and Amazon Alexa may allow their users to trigger an interaction, the experience is neither implicit or collaborative: the user must utter a specific, fixed key phrase to interact with the system [106], and while the system can learn and update its configurations from the user's speech-to-text

inputs, the degree of such changes remains opaque to end-users and subject to legal scrutiny surrounding data privacy and the Right to be Forgotten [156]. Custom AI systems are free from such controversies, yet they are subject to another limitation: while these systems allow the architects to implicitly monitor user behavior and eagerly collect their insights and actions, traditional machine learning models often require offline training based on a snapshot of accumulated data, preventing true real-time decision making by AI systems [116]. An ideal system should allow for both flexible and accurate system triggers and establish a more complete "feedback loop" between AI and human users, or better yet, a truly collaborative experience where the two parties can work in parallel in mixed-initiative tasks.

The "black box" nature of conventional AI systems has been a subject of both mystique and notoriety [64], but it is especially more detrimental to facilitating human-AI collaboration and erodes human users' trust and reliance on machine teammates [122]. In addition to improving performance and ensuring reliable operation, AI systems can potentially combat mistrust by making their inner workings visible to human users and help them feel at ease with transparency. The benefits of such transparency over AI systems' decision-making processes are well documented and thoroughly discussed [30], yet conventional AI systems offer little more than complex network visualization of underlying AI models, conducive to debugging rather than understanding [134]. While exposing the underlying algorithm is equally valuable to improving system performance, there are opportunities for AI systems to gain the trust of human users and make them understand how they operate [24].

2.3 Challenges in Designing Collaboration

Beyond the improvements made to communicative AI systems, the very ways of interaction and collaboration between the two parties are subject to scrutiny as well. In an ideal collaborative setting, one must be able to introduce new constraints without radically changing the entire plan, while being able to flexibly adopt innovative ideas and shift strategies for problem solving. Yet, tra-

ditional planning representations and algorithms are less suitable in "incremental, user-centered collaboration": after all, a human planner with access to rigid blueprints will not be able to incrementally update objectives, change constraints, and suggest partial solutions as the situation develops [3]. Several challenges emerge as the focus shifts from the AI agent itself to collaborative settings that involve human users.

Similar to the way human-only organizations operate, future human-machine teams could be assembled based on specialties and capabilities of individual teammates. In addition to optimizing teams based on competency, some imagine machine teammates that may rise to the role of a team leader beyond active participation [143]. Meanwhile, an invisible tension between human users and AI agents remains as machine teammates slowly assume tasks once considered exclusive to humans [39, 145], and a compelling question surrounding the division of labor emerges [70]. Despite these existentialist concerns, however, today's human-AI collaboration remains a cost- and labour-intensive endeavor that involves human facilitators who help to bridge the gap between domain expert practitioners and novel AI systems but also serve as a key bottleneck in the overall collaborative process [16]. Efforts to automate away such facilitation unfortunately remain expensive [130], and domain experts without collaboration experience face an increased chance to fail. The current state of human-AI collaboration presents an opportunity to further eliminate redundancies and allow the core members to work together with minimal human intervention.

In addition to challenges in forming an effective human-AI team, individual tasks that each party is responsible for should also be carefully optimized for collaboration. A creative practice, such as composing music or producing a visual work of art, may be decomposed into a series of tasks that may be delegated to the AI system, and such AI agents may be embraced by artists who are "open to working in new ways" [34]. On the other hand, some high-stake collaboration scenarios may result in negative user experiences, where the AI partner may produce results at odds with the human user's prior expert knowledge [8]. In this amorphous area, there is a movement to construct a structured taxonomy of tasks suitable for "hybrid intelligence" [32], all while determining the answer

to the questions of "can we" and "should we" by determining delegability of each task based on motivation, difficulty, risk, and trust [95]. Meanwhile, each case for human-AI collaboration continues to be evaluated on a task-by-task basis by human engineers and expert facilitators [130].

More than so human counterparts, machine teammates can be trained to specialize in specific collaboration processes, including "coordination, knowledge sharing, or evaluation" [143], and such processes can be further influenced by the mode of communication shared by human users and AI agents. As consumers interact with their smart speakers with voice, musicians may interact with their AI assistants using MIDI-enabled digital instruments; while business intelligence applications may rely on various explicit mouse clicks and keyboard shortcuts, other camera-activated AI solutions may watch for specific poses by their human partners. Such explicit interaction modes can also be supplemented with nonverbal cues: with human users particularly sensitive to unrelated utterances and expressive gestures [88], their AI assistants should also be able to "return the favor"[15] by reading extra cues that "expose each human user's mental and emotional states"[102].

2.4 Barriers to Widespread Adoption

Upon further defining technical and contextual aspects of human-AI communication, one must turn attention to the way such technologies are governed and adopted by the society at large. As AI products become more readily available, ethical and moral challenges emerge, with questions surrounding "unintended consequences that threaten human autonomy" as well as the moral code that each AI agent is expected to follow [143]. Human dynamics of trust and reliance also play a role: while adults are capable of resisting simulated social pressure and ignore incorrect AI recommendations, children are susceptible to "caving in" and simply following the AI agent's lead [158]; in some cases, human experts choose to "act defensively and ignore the systems conflicting recommendations," indicating that they may be feeling threatened by the efficacy of AI assistants and display mistrust [41]. Such social and institutional barriers may require

specialized regulations and organizational norms amongst the human-AI teams.

Machine teammates are seemingly exempt from issues of responsibility and liability, leaving such policy-related discussions to human stakeholders. In order to build comprehensive AI regulation that does not stifle innovation, researchers and regulators need to work together to bring clarity to the right and the obligations of individual stakeholders and build confidence and acceptance of AI assistance. Researchers and designers must ask themselves what decisions can and should be deferred to AI agents, and ruminate whether they will behave in the best interests of human partners [115]; lawmakers must support this process by "codify(ing) overarching principles" and "adopt industry and market-specific legislation" as certain AI systems approach maturity [19].

As the regulators and the industry proceed with building a sustainable infrastructure to support human-AI collaboration, human users will need to learn to adapt and embrace their AI partners. In business settings, leaders should "convey the urgency of AI initiatives" and "invest in AI education for everyone" [47], training in required competencies for collaboration. Meanwhile, schools are recommended to adopt "a staged approach" through the AI journey as part of the required curriculum [119].

2.5 Summary

Recognizing the surge of interest in utilizing AI systems for joint activity, this chapter identified a number of barriers preventing the widespread adoption of human-AI collaboration. These challenges also suggest a number of opportunities that this dissertation sets out to further explore in facilitating future implementation. This chapter can be summarized as the following:

- Conventional AI systems are limited in their ability to customize appearance, learn human insights in real time, and explain their actions.
- Team composition, task design, and mode of communication need to be tailored specifically to collaborative scenarios.

- Human stakeholders should work together to establish a sustainable socio-legal framework conducive to human-AI collaboration and prepare business and education sectors with the necessary training.

Chapter 3

Modular Interface Framework

What makes a mechanism is the separation and extension of separate parts of our body as hand, arm, feet, in pen, hammer, wheel. And the mechanization of a task is done by segmentation of each part of an action in a series of uniform, repeatable, and movable parts. The exact opposite characterizes cybernation (or automation), which has been described as a way of thinking, as much as a way of doing. Instead of being concerned with separate machines, cybernation looks at the production problem as an integrated system of information handling.

Marshall McLuhan, *Understanding Media: The Extensions of Man* [103]

3.1 Introduction

To simulate a certain human ability using artificial intelligence is to understand the surrounding context, dissect the corresponding activities into repeatable processes, and train an AI agent to reliably do the same. Regardless of their outcome, such endeavors truly feel as labour- and time-intensive as carving a bronze sculpture of a human figure or conducting a nationwide census — requiring careful dissection, collection, and analysis of human condition.

These monumental challenges, however, are made less daunting with the

rise of crowdsourcing platforms. Some, inspired by financial incentives, flock to Amazon Mechanical Turk platform to complete individual HITs (human intelligence tasks), ranging from answering simple surveys to labelling a complex photograph [66]; many solve difficult reCAPTCHA challenges to sign up for a website account, identifying various distorted English words, without realizing they are digitizing New York Times archives one word at a time [159]; a few dedicated individuals may leave their networked computers overnight and "donate" their computing resources to various scientific projects, such as Folding@Home [12]. Whether driven by tangible rewards or pure curiosity, distributed efforts yielded significant progress towards large-scale datasets: Mechanical Turk continues to serve as a popular method for researchers to generate crowdsourced datasets with [21], and such datasets feature named entities [83], and behavioural patterns [133], and emotive words and phrases [111], all contributed by people around the world.

The excitement continues to build as machine learning applications integrate data beyond simple texts and numbers, such as language and vision. Ranging from simple captioning of static images to describing scenes in films, the rising interest in such research areas has also yielded in a plethora of new datasets released to the public, as well as an appetite for many more [46]. Such datasets are designed to highlight visual and textual correspondence, context, and narratives, and generating one understandably requires a lot of work: each research group ends up building a bespoke tool to collect and organize annotations [123]. While web annotation tools and services do exist for public access, such tools are limited to specific modalities and do not offer joint annotation of text and vision data. This gap in the market results in a series of fragmented and costly efforts by these research groups, often prone to initial defects and multiple iterations common in a software development cycle, as well as difficulty in large-scale deployment [94].

There also exists a gap between tools for data annotation and interfaces of machine learning systems, and understandably so, as they are created for two distinct user groups: annotators who fuel machine learning projects, and end users of such systems. However, with both tools sharing a large set of

interface components and workflows, there lies an opportunity to offer a common approach to data annotation and visualization — with applications to artificial intelligence and machine learning.

Recognizing these present gaps between dataset annotation interfaces and machine learning visualization solutions, this chapter presents Modular: a modular annotation and visualization framework and the accompanying proof-of-concept software platform that enables researchers to rapidly set up an interface for annotating new datasets and visualizing predictions made by a machine learning model. The platform enables many of the standard and popular visual and textual modalities available in conventional annotation tools, configurable not only to collect data, but to visualize outputs of existing machine learning systems and even launch hybrid initiatives such as user studies. Finally, as the name Modular entails, research engineers and scientists can also extend existing modules or create entirely new ones that are specific to project needs — optimizing the workflow for researchers and enabling them to seamlessly conduct their work without being bound by out-of-the-box solutions or building custom software.

3.2 Related Work

Annotation tools currently available for public access exist along a spectrum between the two distinct types of annotation: text and pixel (for static and moving images). While the majority of tools specialize in a specific type of annotation, a handful of multimodal tools set out to allow their users to create different types of annotation and bind them together with mixed results.

3.2.1 Text Annotation Tools

Ranging from simple free-form text entries to structured word-based tags, text annotations come in a variety of forms, and there exist numerous tools designed for domain-specific use cases. BRAT [150] provides features for structured annotations with fixed-form text, where the users can mark specific tokens with text labels and color blocks and connect them with simple associations. Meanwhile,

Webanno [36] provides a focused set of linguistic features with a multi-user interface, allowing a group of users to collaborate on a larger body of text using a strictly defined set of morphological, syntactical, and semantic annotations. Finally, Knowtator [117] offers an ability to define custom ontologies powered by the popular, open-source framework Protégé, enabling domain-specific annotation tasks with hand-crafted ontologies.

Optimized for use cases that require collaborative efforts in dissecting large text corpora and handling knowledge management, such text-oriented tools may be a good fit for a range of research domains: BRAT was extensively used in epigenetics and infectious diseases subdomains of biology, while Webanno and Knowtator claim flexible application to different domains due to their standardized knowledge representation. They, however, present some barrier to wider adoption: limited feature extensibility by end-users, steep learning curve, and difficulty in deployment for non-expert access. Often assuming interface designs and workflows conducive to specific problem domains, these tools are rigidly purpose-built and resistant to extending tool capabilities or making developer- or researcher-initiated interface updates. The result is a suite of tools that require extensive training prior to use and prevent novice users from participating in establishing a collective knowledge base [16].

3.2.2 Pixel Annotation Tools

Necessitated by the emergence of products that rely on AI-based image recognition and generation, there are also multiple open-source tools designed for collecting image and video annotations. Image annotation tools such as LabelMe [138] and Annotorious [148] offer a standardized interface where users can select a single image and create one or more polygonal overlay elements that correspond to different parts of the image, complete with basic labels.

Video annotation presents a challenge beyond labeling different parts of a single static image, as each video file features an overwhelming number of static frame images with multiple entities actively entering and exiting a scene. Tools such as LabelMe Video [167] and VATIC [160] adopt affordances originally demonstrated in static image annotation tools, including bounding boxes and

polygonal annotations with class labels and attributes, and implement features common in video editing software: basic keyframe interpolation and on-screen motion tracking. Finally, CVAT [91], a platform created by the team behind the open source computer vision library OpenCV, offers an assisted experience made possible with state-of-the-art algorithms [97] — allowing users to more quickly and conveniently annotate with automatic entity detection.

Unlike their text-specific, self-hosted counterparts that largely rely on local server storage, these pixel annotation tools fully support cloud storage and dedicated servers to facilitate larger media assets and access for wider audiences. Their user interface offerings remain largely opinionated, however, in terms of how the user's annotation experience should unfold, offering little flexibility in tailoring the user experience and satisfying project-specific needs. While such conventions and practices are designed to facilitate the project at hand, this can contribute to excessive dependence on specific tools with little to no alternatives down the line: this potential issue is well-aligned with the software industry, where the use of system-specific conventions render "the semantics of the system inseparable from the tools" [107] and become a point of contention.

3.2.3 Multimodal Annotation Tools

As the demand for more human-annotated data becomes more complex, the tasks designed to collect such data transform to be more layered and nuanced: an annotator may be asked to transcribe an audio recording and highlight relevant keywords, while another may be requested to connect different characters in an excerpt from a film script to on-screen entities in static images.

While many researchers may "duct-tape" existing single-purpose tools together [42] and request users to annotate the same asset multiple times using different methods [147], other bespoke tools set out to support multiple annotation modes: ELAN [82] allows users to create free-form text annotations for a specific audio or video recording, while NOVA [161] similarly supports text annotations for audio-video recordings and other non-verbal communication cues, such as facial expressions and gestures, represented as continuous datasets. However, while these tools do support multiple types of assets, resul-

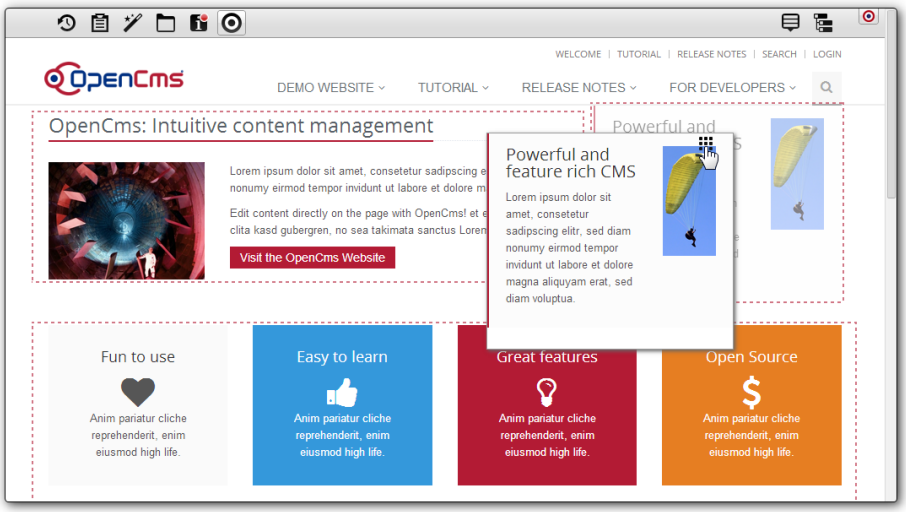

Figure 3.1: When using a conventional website builder, each section of the page is represented as a module, available for the user to customize, populate, and place as necessary [17].

tant annotations remain separate and disconnected, other than being aligned on a timeline. The generically named Universal Data Tool [68], on the other hand, features a large set of annotation methods designed for specific scenarios, ranging from audio transcription to video segmentation, but forces the users to dedicate a project to a specific annotation type without the ability to build custom scenarios or switch between existing scenarios on-the-fly.

While inspired by common annotation scenarios, these multimodal tools fail to allow users to jointly create text and pixel annotations and assign contextual links between individual annotations. Each of these tools addresses a very specific line of research and thus cannot serve projects that cut across computational language and vision disciplines, leaving researchers to their own devices in producing more complex datasets.

3.3 Approach

Inspired by conventional wisdom in building contemporary web-based software, Modular is built on three main approaches in response to the apparent gaps in data-hungry areas in AI research and annotation tools: (1) straightforward

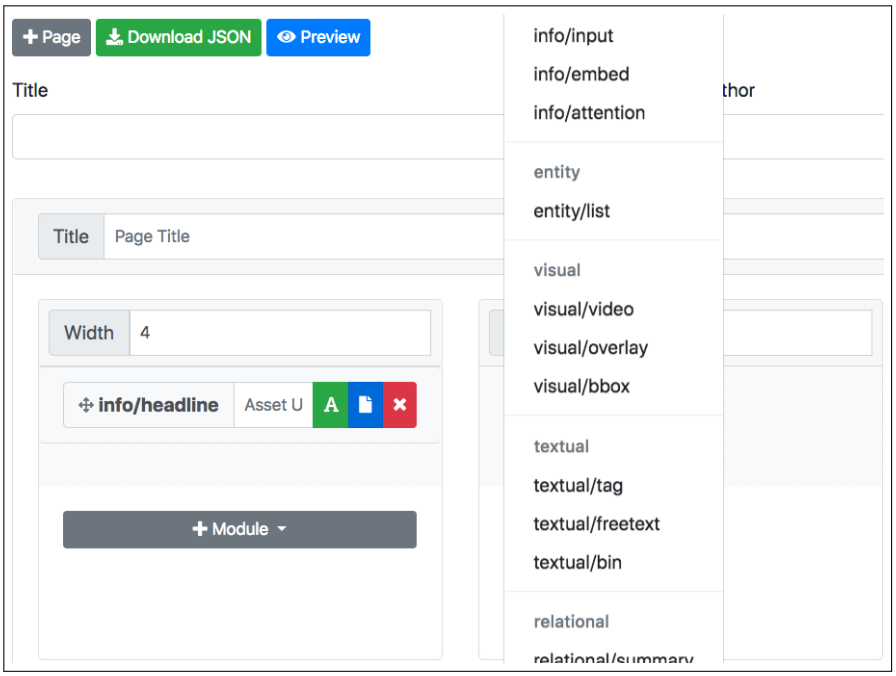

Figure 3.2: Each interface can be easily constructed and customized using Modular's web-based layout generator, eliminating the need for ad hoc web development and easing the initial learning curve. The dropdown menu, triggered by the "new module" button, features all available modules.

user experience for stakeholders on both sides of the interface, (2) reusable, configurable, and context-sensitive modules that respond to changing needs of the users, and (3) ability to extend the functionality of the tool with project-specific code.

Usability

Deploying personal websites is no longer a feat exclusive to programmers or dedicated hobbyists, as today's website building tools now allow a much larger demographic of users to customize and publish websites with ease. As illustrated in Figure 3.1, people can click and drag customizable elements, ranging from text boxes to video players, into an empty canvas, and one can further arrange the modules and insert additional editorial content as required.

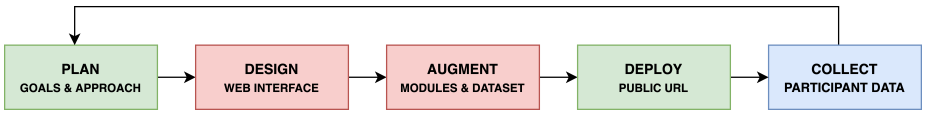

Figure 3.3: In this representation of a typical user experience in designing and deploying an interface using Modular, the researcher (green) can define and communicate an overarching goal. This is then translated to a web-based experience by the designer-developer (red) using a drag-and-drop interface, coupled with reusable modules, custom implementations, and applicable datasets. The resultant interface is then accessed by the user (blue) using a unique URL, and the collected data is delivered to the researcher.

Modular sets out to apply the above paradigm to collecting annotation data and visualizing AI models. While each iteration of its framework can be manually constructed using a plaintext layout specification file, Modular's web-based layout generator sets out to ease the burden of coding for typical lay-users by emulating a WYSIWIG ("What You See Is What You Get") approach found in popular website builders such as Wix and Squarespace. As illustrated in Figure 3.2, the layout generator allows the user to customize the placement and the size of individual modules, and also dictate how the whole experience unfolds using pagination. Using the interface, each researcher can conveniently create a simple annotation tool, visualize datasets, or deploy a complex user study — and the participants on the other side can easily access with a "permalink," eliminating the need to install dedicated software. A typical user experience of using Modular is illustrated in Figure 3.3.

Contextuality

In out-of-the-box tools and other research-specific efforts, each interface component is often rigidly configured to perform one or more specific tasks: a text tagging component may be dedicated to receive text inputs from the user, necessitating a separate component to visualize the very same data collected from other users. The results are costly redundancies and a lost opportunity in building reusable components whose visual elements and behavior transform based on the context in which they are used.

Pursuing context-aware behavior with minimal redundancy, all modules of-

fered by Modular are designed to perform dual functions: to collect annotations (write) or to visualize a provided dataset (read), which can be toggled when establishing the layout.

This behavior is further segmented and defined to vary based on co-location. For instance, if an instance of a text tagging module, set to annotation mode, exists alongside a bounding box module configured to annotate another image in identifying notable characters, another module — namely a graph module — can be configured to treat all annotations, textual or visual, as viable nodes that can be connected to one another. In more granular cases, individual events emitted by one module can trigger a response in another module: while the user "scrubs" a video looking for a specific character, a list of pertinent characters in another module may light up based on the video player's temporal data. These relationships and behaviours have been manually identified and established at the framework level, but they can be disabled or reconfigured as per project needs.

As illustrated in Table 3.1, Modular presents an opportunity where the same modules can come together perform different functions based on their configuration and placement, eliminating the need to build a wholly separate yet largely redundant component each time a new project requirement emerges.

Extensibility

The various modules offered by the framework are designed to sufficiently cover common use cases explored by other tools, ranging from tagging images to conversing with AI agents, yet there is a use case unique to a research project that necessitates a level of customization. Consisting of standard page, style,

	Annotation	**Visualization**	**User Studies**
Text	Word Token Tags	Language Parser	Freetext Entry
Video	Landmark Points	Entity Detector Overlay	AI-generated Clips
Graph	Text-Visual Links	Spatio-Temporal Links	Audience Clustering

Table 3.1: Sample manifestation of interface modules applied to various research efforts, including conversational AI, explainable AI, and commonsense grounding.

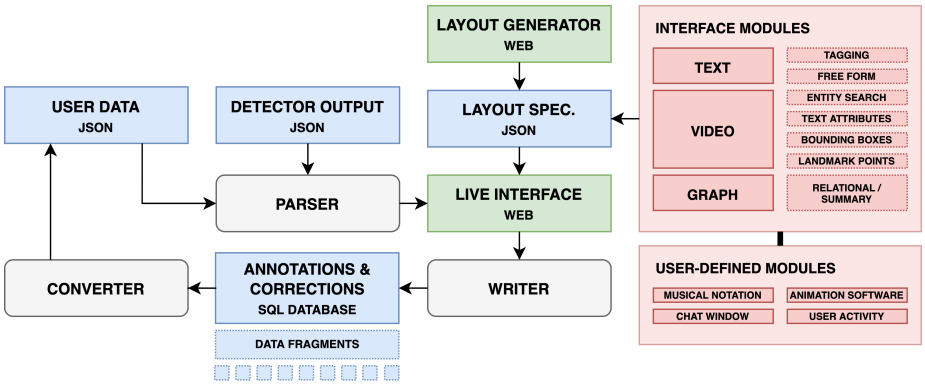

Figure 3.4: The interface consists of text, video, and graph modules, which are arranged as a plaintext specification file. The modules interact with a backend that loads the data schema and saves and loads annotations, which can be exported as a single output file. Independent by default, inserted modules can be manually configured to share the same data model and behave as a single unit. The user can also create and inject new project-specific modules as necessary.

and behaviour files, all existing modules follow the same construction, and other users of Modular may freely extend existing modules or create entirely new modules per project needs.

3.4 Interface Modules

Modular's interface hosts a number of independent, context-free components that can be freely arranged and combined as the research or dataset need emerges. Summarized in Figure 3.4, they are organized into three distinct categories: text, video/image, and graph. Inspired by existing single-type and multimodal annotation tools, these modules offer a variety of ways to annotate a dataset or visualize results. The designer may create custom modules specific to project needs, such as audio, geolocation, and spreadsheets.

3.4.1 Text Module

Consisting of simple word token tagging and free-form text input components, the text module enables each user to mark different terms or create comments

pertaining to a specific part of a text-based dataset. These user-provided annotations can also introduce an element of collaboration (or contention) as the interface may offer a visual indication of previously tagged terms, as well as a visualization of agreement amongst uploaded entries as deemed necessary by the developer. These features are further described in subsequent sections.

Tagging

The tagging component, illustrated in Figure 3.5, tokenizes individual words, paragraphs, and sentences to enable token-based tagging across the interface. Upon discovering a notable word or phrase in the dataset, the user can click to select one or more individual tokens. The component also captures other types of metadata such as time stamps or presence of other interface modules, as manually configured in the layout generator, for persistent storage. The module also enables colored tokens to allow for importance annotation or visualization of output from probabilistic models producing importance weights per token, such as models that feature learned attention.

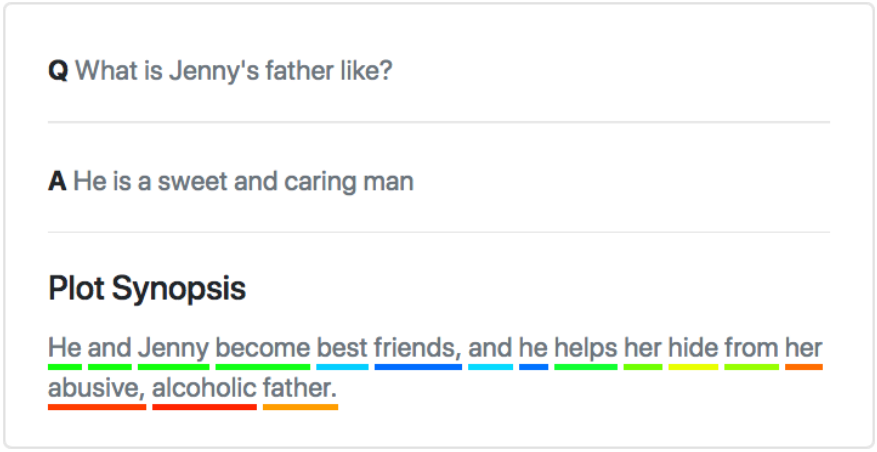

Figure 3.5: The tagging module (top) enables marking specific words. The user is asked to inspect the question-and-answer pair, along with a series of relevant word tokens and visualized importance weights based on the dataset injected by the developer. The user may flag individual tokens pertinent to the question with mouse click: flagged tokens, now in inverted color to indicate selection, are communicated back to the researcher for future reference.

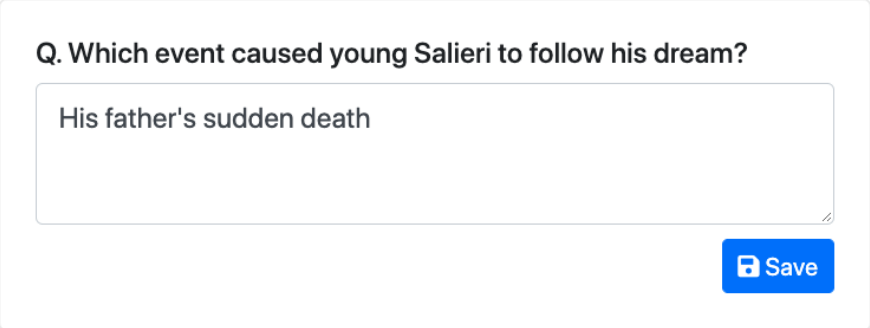

Figure 3.6: The free-form module (bottom) allows the user to submit plaintext data, whether to answer specific questions or record observations.

Free-Form

Also illustrated in Figure 3.6, this component allows the user to submit free-form text entries that further annotate or describe the dataset. Each submission serves as an accompanying annotation to token tags or as a standalone comment, and contains the same set of comprehensive metadata as token tags.

3.4.2 Video Module

The video module features a full-motion video player with a set of interactive overlay components, informed by existing video annotation tools and available datasets. The user can activate each component to reveal more insights pertaining to the video, and further interact with individual entities to insert annotations or augment the original dataset.

Bounding Boxes

Visual entities, ranging from algorithmically detected objects to manually annotated counterparts, can be represented as 2D bounding boxes illustrated in Figure 3.7. Each box displays above the player component, moving in real-time along with the video. Based on the parameters provided by the developer, each box may be marked in a different color, display in a varying opacity, or contain text-based labels such as attributes. Using the module, the user can also directly interact with bounding boxes to adjust their size or position, edit their

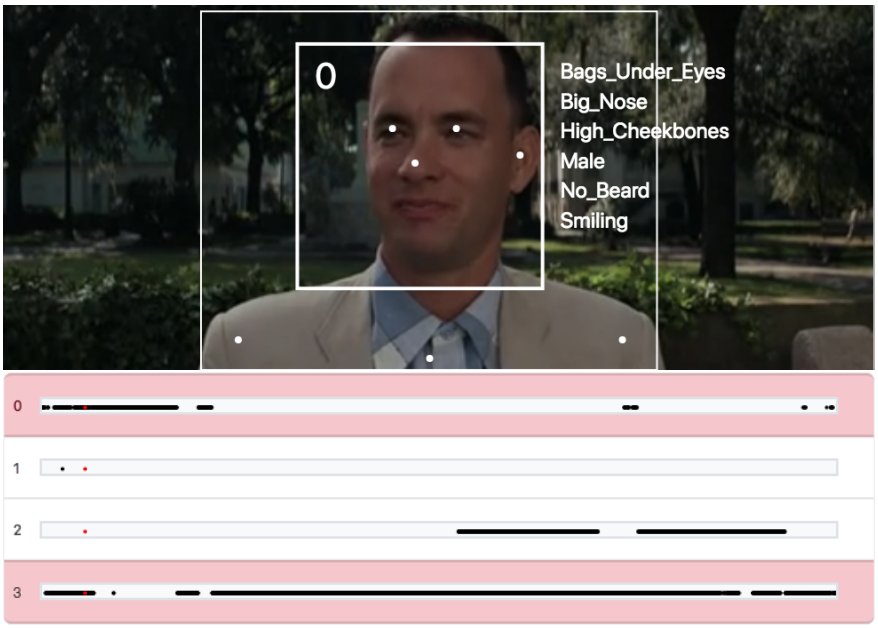

Figure 3.7: The video module (top) with the entity seek module (bottom) displaying different entities present in the video and their temporal annotation. Each bar in the entity seek module indicates the presence of individual entities, marked by their index numbers, across the timeline.

labels, or create additional boxes as required.

Polygons / Segmentation

The user can also create and modify polygonal annotations. Taking cue from popular graphics editor applications, the component allows the user to click on specific parts of the frame to construct a polygon and place it upon an object. The generated polygon serves the same function as bounding boxes, including spatial interpolation between keyframes.

Landmark Points

The video module also offers the ability to visualize a group of anchor points over video, as illustrated in Figure 3.7. Suitable for representing skeletal and facial feature tracking data, the resultant overlay dots can also vary in their

opacity, size, and color as per user-specified parameters, and are also available for user modifications and annotations. These visual attributes can be used to visualize different dataset properties, such as confidence scores and entity identifiers, as applicable.

Text Attributes

In addition to visual entities that display (and move) in synchronization with on-screen objects, text-based overlay options are available for the user as illustrated in Figure 3.7. Text-based subtitles and captions coincide with dialogues in the video, and offer the user the same degrees of interaction as the token tagging component found in the text module: the user can click on one or more individual word tokens to simply mark as notable or annotate with free form text comments.

Entity Seek

This module allows the user to look for occurrences of a certain entity across the video timeline illustrated in Figure 3.7. Characterized by a timeline visualization situated below the video progress bar, the feature displays an on-screen or in-script presence of an entity with a series of dots, indicating that the user can "scrub" the video to a specific point of the timeline to discover that particular entity. As the individual timeline components "light up" when the corresponding entities are displayed on-screen, the module also allows the user to easily identify co-occurrence of two or more distinct entities.

3.4.3 Graph Module

Many annotation tools are designed to support a limited set of media files and annotation types, and a disconnect between tool design and user needs quickly emerges as the tool starts to support flexibility [31]: to bridge the gap between different types of annotation, Modular introduces a unique graph module. This module serves as a versatile method of generating 2D node-link graphs in direct relation to on-screen visual entities and/or text token nodes. Each generated

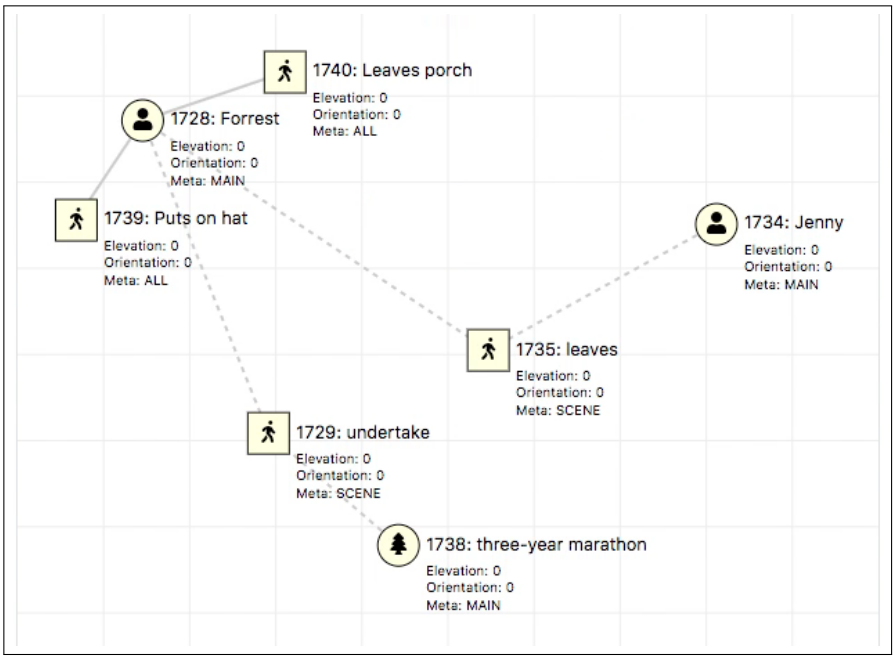

Figure 3.8: A 2D graph visualizing the relationship of different on-screen entities and text tokens using nodes, icons, and edges. Initially created using force-directed graph drawing, the graph can be manually modified by the user with click and drag.

graph can be presented as a single static image or a spatio-temporal animation displayed in synchronization with video playback.

Utilizing SVG, the popular vector graphics format, the two-dimensional graph component features an ability to generate vector-based network graphs that visualize relationships between different entities, including individual on-screen visual tokens and text tokens, in the dataset illustrated in Figure 3.8.

Each entity is represented as a node, with its various visual attributes, including size, opacity, color, and icon, mapped to user-defined characteristics of the corresponding entity. Two or more nodes may be linked using one or more path objects, each equipped with its own set of modifiers: path type (dotted, solid), direction (bidirectional, unidirectional, or non-directional), and a text-based annotation.

Finally, the resultant graph can visualize hierarchical information by using a tree-like approach: each of the larger, main nodes can have a series of child

nodes, which in turn have the capacity to have their own set of child nodes. While each node in the graph can be placed randomly in the canvas, the user may toggle a trigger that maps the position of each node to the corresponding entity in a video clip allowing the graph to capture the on-screen spatio-temporal information as well: for instance, a graph node representing a main character may move in accordance with the character's on-screen movement across the scene.

Instead of simply inspecting the resultant graph in a passive manner, the user may actively interact with nodes and links to induce changes to the dataset. The user may click and drag a child node and simply migrate it from one parent node to another in order to swap the two entities' characteristics at the dataset level. Alternatively, the user may remove a link between two nodes to sever the relationship between the two entities, or reverse the direction of the inter-node link to update the nature of the relationship.

3.4.4 User-Defined Module

Beyond the original offerings of the framework, the developer may wish to define entirely new experiences specific to project needs, ranging from modules that provide data summary based on previously created user annotation to those designed to allow users to directly interact with external AI agents or solutions. Below are some of the user-defined modules necessitated by the collaborative work discussed in the subsequent chapters.

Summary

With all user-provided annotations stored in a database table, the module embraces and promotes collaboration and collective efforts by visualizing the summary of prior user activities. Illustrated in Figure 3.9, the summary module also may integrate with the previously described graph module to generate a simple network visualization of relationships among the user annotations, and offer a download link that allows the user to download all the accumulated annotations. This module was originally constructed for a prototype of a crowdsourced annotation project, designed to allow the administrators to have a consolidated

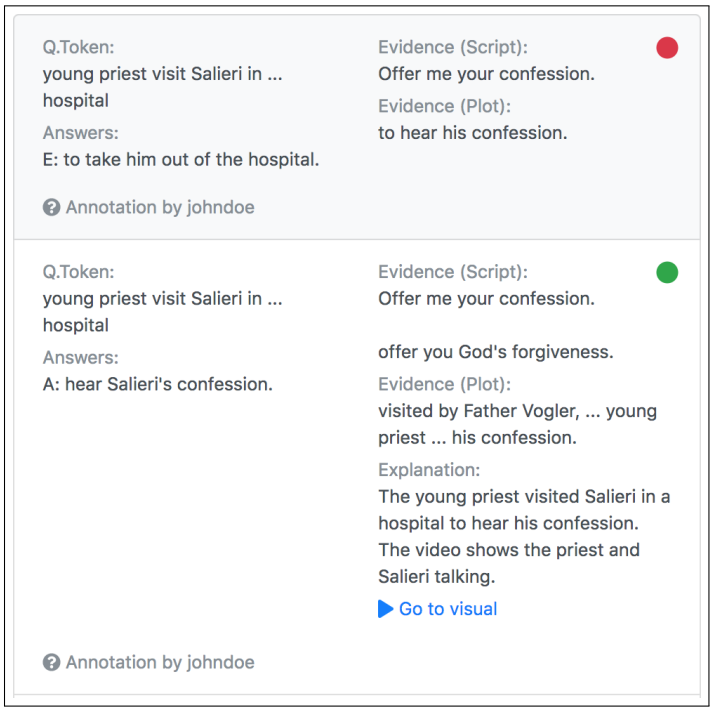

Figure 3.9: An example of a user-defined summary module, designed to display a list of annotations created by different annotators.

view of annotations created by a large number of anonymous users.

Activity

Upon entering the interface, the user can immediately identify the in-text keywords that have been annotated or tagged by other participants. Each previously-tagged token is displayed with varying visual attributes as inspired by micro-encoding techniques [14] — font weight, opacity or color saturation — alerting the user to how popular (or unpopular) that token is. The user can then click on the token to reveal all submissions associated with the term illustrated in Figure 3.10. The existing video player module may be extended to visualize the level of user activity across the timeline, indicating more popular (or contentious) points of each clip. The opacity or saturation of each color block represents varying degrees of user activity at a more granular level.

Figure 3.10: An example of a user-defined activity module, where green indicates entries that support a specific question-answer pair, while red indicates those that serve as refuting evidence. The developer may highlight individual word tokens to accommodate stylistic needs or visualize user activity surrounding them. The underline below the token indicates selection state, and upon click, the user can further explore detailed activities surrounding the selected token.

Agreement

This module also provides a more thorough visualization of user submissions, clustered by their response categories. Each bar, illustrated in Figure 3.10, below the corresponding statement displays the makeup of associated responses, allowing users to quickly identify the popular opinion (or lack thereof) pertaining to each statement. Like the previously mentioned Summary module, this module too was created to reduce the visual and cognitive load that may be imposed on the administrator in reviewing raw database table content.

Conversation

Featuring a modality similar to today's messenger applications, the conversation module, illustrated in Figure 3.11 allows the user to directly engage with an external solution using a chat-style interface. Whether it be an AI model featuring

its own natural language parser or another web service pertinent to project requirements, the external solution processes and returns an output. This output take a form of a simple text message or a complex payload, which can simply be printed by the conversation module or interpreted by other modules that can take advantage of it: this module was especially useful for projects that support text-based collaboration between the human user and the AI system, described in the future chapters.

3.5 Back End

Built with simplicity and extensibility in mind, Modular relies on a series of popular core web technologies with little dependence on niche plugins. The Bootstrap framework serves as a basis to the various interface modules, while PHP and MySQL serve as backend supports that handle record storage and re-

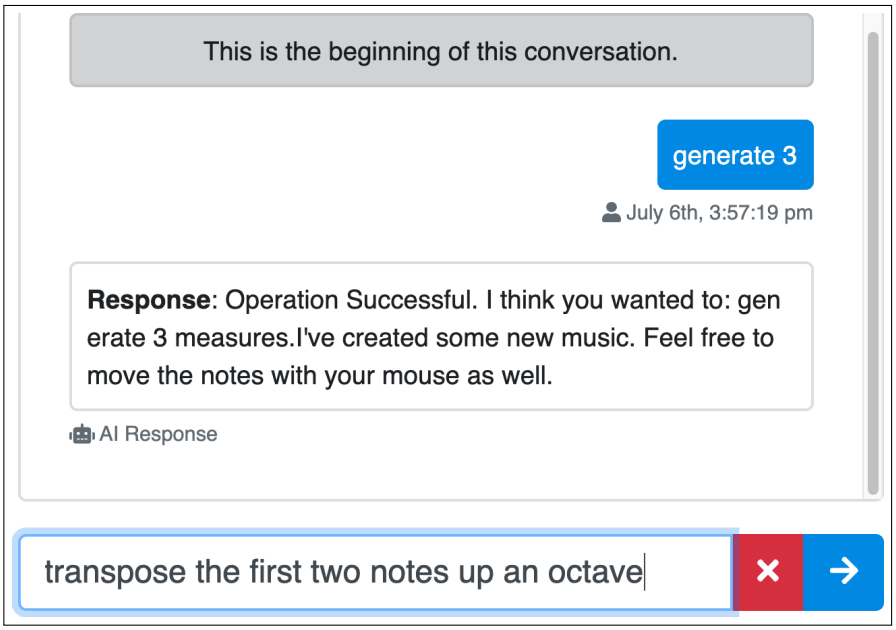

Figure 3.11: A conversation module, where the user can provide a text input to directly engage an external AI model or web service. This module can simply display the returned text message, or trigger other modules to react to the available output.

trieval. Finally, JavaScript and jQuery serve an instrumental role in integrating the various interface modules into a cohesive experience. Separation between visual elements, presentation style, and behavior present in individual modules is inspired by precedents established by these technologies.

General Data Structure

The dataset file structure is designed to support weighted text tokens, free-form text, bounding boxes, text attributes, and landmark points. Weighted text tokens consist of the token identifier along with the associated weight. In the case of unweighted tokens, the value of the associated weight is simply set to "null". Associated free-form text is saved directly as a string.

Bounding boxes consist of the coordinates of the box's location, width and height, and an object label, a confidence score (automatically generated by a machine learning model), and a tracked identifier derived from any pertinent AI model. Finally, landmark points for facial or body pose comprise a list of 2D coordinates along with the confidence value for each point (provided by a pose estimation algorithm), tracked identity, and other miscellaneous attributes.

Database

While the user is expected to directly manipulate the visible data using the various interface modules, the underlying dataset remains intact without any permanent, irreversible changes. Instead, the interface pushes an incremental change, or a "delta," to the database table in order to maintain version control and activity history.

Each delta entry consists of three main components: the timestamp, the session identifier (which doubles as a username), and the text fragment designed to replace the original counterpart. Upon detecting the user's interaction with the applicable entity, whether it be a text token or a bounding box, the interface creates a copy of the underlying data object's schema. This schema is then populated with the user-generated values, and then committed to the table.

When initializing the interface, the database module loads all the incremental changes and the original dataset into the memory. Upon completion, the

interface proceeds with merging the two datasets by "injecting" each relevant delta into the dataset, producing a merged version for the interface to reference. This process takes place regularly in the background as the user continues to make corrections and annotations.

Version Control and Data Exporter-Loader

Similar to version control systems such as Subversion or Git, this feature allows the user to identify the differences and revert to previous annotations or original data and download all the accumulated annotations, along with the schema supporting them. Once the user acquires the file, one can load it back to the interface to visualize, modify or augment.

Module Structure

Each module consists of visual element (HTML), presentation style (CSS), and behavior files (JavaScript). Upon passing a series of basic syntax validations to prevent compilation error, each module can be readily added to the user-facing interface via the layout generator. Element and style files work to define the look-and-feel of an individual module, while the more complex behavior file is responsible for establishing the available actions in each module, as well as its behavior in relation to other co-located modules and their own events.

For example, the developer may wish to build a custom module that presents visual interestingness of an image, and ensure that this module refers to the image content of the present video player. Upon establishing the visual elements of the module, the user can instruct the module to look for the video player module, capture the still image, and analyze its image content. This particular custom behavior can then be "scheduled" to trigger every time the player moves to the next video frame.

3.6 Use Cases

Described in Table 3.1, this framework can applied to a number of different aspects in interdisciplinary research efforts across visualization, annotation, and

user studies.

Visualization

The user can build an interactive visualization of available datasets or results without resorting to creating bespoke software. After constructing the layout and inserting the necessary modules, the researcher can attach a large dataset, a plain text file, or a video file to each relevant module prior to deployment. The resultant interface will automatically load the assets as specified in the modules. These aspects of the framework are thoroughly explored in subsequent chapters.

Annotation

Beyond passive visualizations, the user can actively interact with the modules and create new annotations to build a new dataset or contribute to an existing one. The user can watch a video, identify a series of on-screen entities, and create a series of bounding boxes. All the activities are recorded and become available for download as a plaintext file. A large-scale annotation project, designed to synchronize multiple discrete datasets and allow human users to further augment the dataset, described in Chapter 4 relies on this use case.

User Study

The framework also supports a lengthier, more complex experience where the user is guided through a series of different annotation, inference, and analysis tasks. Spanning multiple pages and equipped with a variety of editorial content, the interface presents an opportunity for deploying large-scale user studies without deep technical knowledge. Chapter 7 relies on this use case to record, observe, assess trust and reliance in human-AI collaboration.

Target Users

Modular is designed to facilitate planning, prototyping, and deployment processes that often involve multiple stakeholder groups from a larger organization. The framework, by design, allows all stakeholders — researchers, developers,

designers, and other curious enthusiasts — to independently oversee the end-to-end process of building a user-oriented web experience with minimal disruption as the provided modules are sufficient for basic projects. However, using Modular in a highly collaborative environment will most likely result in a positive experience as all stakeholders can contribute to further augmenting the "vanilla" Modular framework and tailor to project needs.

Commercial License Agreement

While Modular is freely available for academic and non-commercial research use, the framework has been commercially licensed to SRI International, whose resident researchers served as close collaborators for projects introduced in the subsequent chapters, with Ontario Tech University serving as the Intellectual Property agent. Modular remains in use for SRI International's projects beyond those discussed in the dissertation.

3.7 Summary

Having identified opportunities for improvement in tools currently available for data annotation and model visualization in areas of AI research, this chapter introduced Modular, a new interface framework underpinning the projects discussed in the subsequent chapters of the dissertation. This chapter can be summarized as the following:

- Research efforts surrounding data annotation and AI model visualization remain fragmented due to lack of flexibility of existing tools.
- Popular approaches in web development can be applied to deploying annotation and visualization projects in a user-friendly, cost-effective fashion.
- Module-oriented approach to building a user interface can be used to construct user experiences with minimal redundancy and flexible configuration.

Chapter 4

Annotation for Commonsense Grounding

Common sense is not a simple thing. Instead, it is an immense society of hard earned practical ideas — of multitudes of life-learned rules and exceptions, dispositions and tendencies, balances and checks. If common sense is so diverse and intricate, what makes it seem so obvious and natural? This illusion of simplicity comes from losing touch with what happened during infancy, when we formed our first abilities. As each new group of skills matures, we build more layers on top of them. As time goes on the layers below become increasingly remote until, when we try to speak of them in later life, we find ourselves with little more to say than "I don't know."

Marvin Minsky, *The Society of Mind* [109]

4.1 Introduction

As artificial intelligence takes center stage with impressive and even controversial advancements in areas including facial recognition, self-driving vehicles, and even e-sports, a more primal question emerges: can AI view and interact with the everyday world at large in the same fashion as humans do, beyond controlled environments and tightly defined domains?

Despite its name, simulating "common sense" (dubbed commonsense) con-

tinues to be an illusive, distant goal for today's AI research. Whether it be an ability to identify a father in a family photo or attribute a pronoun to a correct object in a sentence [29], commonsense seems well-illustrated with various examples, though lacking a concrete, clear definition in the context of artificial intelligence. Reaching out to the formative years of the AI domain, however, reveals a foundational idea: "we shall therefore say that a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows" [101].

Concerned with the challenge of imitating human understanding of ordinary situations with real-world knowledge, commonsense reasoning is a crucial element in various AI tasks [29]: from disambiguating pronouns in complex sentences to explaining why a character bursts into a laughter, there are many problems that humans can naturally solve with a limited set of clues as we can infer the existence of objects not described in the paragraph and recall the theme of the film to better analyze the scene. Yet, AI continues to struggle with the same challenges, despite the wealth of data collected from people around the world as it struggles to "connect the dots" and "read between the lines" without careful annotation provided by human users.

Scene understanding is a popular challenge in the domain of commonsense: humans, with one glance at an image, can "immediately infer what is happening in the scene beyond what is visually obvious" [170]. On the other hand, AI may be able to identify the characters in the scene with its computer vision model, it certainly could not tell you the motive of the main character on its own. This current challenge, however, presents a unique opportunity where humans can "teach" their AI counterparts how to better interpret a scene and build a cohesive narrative.

In addition to identifying where the characters are standing in the scene and what is being uttered by the main character, AI should be able to tell a coherent, if not compelling, story — and we can enable this shift by contributing more complex annotations beyond narrow text and computer vision data. In doing so, the human partner will also be contributing to commonsense reasoning through the process of "commonsense grounding," where AI gains a better

understanding of the world as a result of this process. After all, a story is more than a sequence of events: storytelling requires multiple types of interactions, such as making logical continuations of the narrative, understanding common story arcs and tropes, understanding non-verbal communication including facial expressions and gestures, and even responding to audience cues. A well-crafted story generated via interactive storytelling must be logically coherent, requiring commonsense knowledge.

Complex and creative tasks are involved dissecting an existing story and transferring the resultant insights to existing or new AI models, and there is a strong need to build a platform that can facilitate human-AI communication and support various annotation types as well as interaction modes. Built in response to such emerging opportunities, this chapter presents the annotation component of Modular-powered Aesop [104, 105]: a new collaborative visual storytelling platform designed to enable bidirectional communication between human user and AI to understand and relate both stakeholders' perception of the world. Designed to clarify implicit human knowledge and collect commonsense insights, Aesop enables users to define and ground complex sentences on visual information, such as videos or animations, in order to create shared understanding and establish commonsense with the AI. These capabilities are powered by novel knowledge graphs, which enable spatio-temporal event representation essential for storytelling.

4.2 Related Work

Aesop's annotation component builds on two distinct areas of research: commonsense grounding and narrative analysis. Prior work in these domains, while providing a strong anchor point for Aesop, presents some of the underexplored opportunities in coherent storytelling and implicit representation of the physical or fictional world.

4.2.1 Commonsense Grounding

Consisting of hard-coded physical and linguistic rules and two-dimensional representation, SHRDLU [164] attempted to ground language in a physical world with symbolic abstraction. Wubble World [62] was later introduced to augment this work initiated by SHRDLU, motivated to help the computer learn language in the same fashion as young children. This work enabled learning language used in a physical world by interactive gameplay with an evolving 3D environment and a multimodal symbolic framework [75], achieved by parsing language and retrieving probabilistic semantic representations from the perceived environment.

Additional work in language-based interactions [76, 121] presented a blocks world platform for building structures. The system combines natural language understanding, planning, and dialogue management, and supports communication about structures where goals are shared between the computer and the human using natural language. However, these platforms are limited to simplistic and rigid environments with communication about abstract structures using explicit instructions and well-defined goals.

New platforms that build on existing video games, such as the Minecraft-based Project Malmo [71], focus on end-to-end learning by solving various tasks in 3D environments, ranging from navigation to collaborative problem solving using language. Similarly, the Quake III Arena-based DeepMind Lab [11] focuses on a maze navigation task and has been extended to incorporate language and learning, via an end-to-end approach which combines reinforcement and unsupervised learning [60]. While these platforms are more advanced and conducive to lower barrier to entry, they are limited by the rigidity caused by boxy, inexpressive worlds and narrowly defined tasks assigned to the agents, offering little in the way of interaction with human users.

4.2.2 Visualization of Stories and Scripts

In the domain of visualizing and interpreting narratives in a scalable fashion, tools such as CARDINAL [98], LISA [140], and CANVAS [72] emerged to assist in authoring and analysis of movie scripts.

CARDINAL visualizes the plots of scripts via a number of views: the textual

script view, a timeline and interaction-centric view, and 2D and 3D previews of the scene itself. While presenting a flow feature to enable scriptwriters visualize their script in a semi-automated way, CARDINAL lacks any notion of shared meaning or continual learning as it follows the handling of directive in a very rigid form with no underlying representation or direct way of interacting with the system to define meaning and streamline visual storytelling. CANVAS [72] is a computer assisted visual authoring tool for synthesizing animations from sparsely-specified narrative events. Unlike CARDINAL, CANVAS produces a series of storyboards rather than a live temporal animation, and provides an interface for authoring and pre-visualizing narratives with AI assistance.

Unlike CARDINAL and CANVAS, LISA and PICA [45] focus on the narrative structure. LISA is an assistive tool for story writers that provides feedback on inconsistencies in the story using artificial intelligence, whereas PICA is a conversational agent for interactive narratives with an underlying knowledge base with encoded belief models for multiple users and autonomous agents in addition to the actual story knowledge. LISA and PICA have some similarities to Aesop, with a single focus on narrative, whether through user interaction or conversation. However, the focus stops at the text level without integration of any visual information.

The above tools and their capabilities are illustrated in Table 4.1, along with the Aesop counterparts. In this comparison, Aesop emerges as the only tool capable of establishing a complex knowledge graph, accepting a variety of media assets, and facilitating both annotation and visualization tasks.

	Knowledge Map	**Media Assets**	**Use Case**
CARDINAL	Entity, action	Script text	Story assistance
CANVAS	None	Storyboard images	Pre-visualization
LISA	Entity, action	Plot text	Content feedback
PICA	Entity, action, event	Chat message text	Knowledge-belief model
Aesop	Entity, action, event	Multimedia	Annotation, visualization

Table 4.1: Comparison between Aesop and other tools for authoring and analysis of movie scripts, in the context of shared knowledge representation, accepted media assets, and primary use cases.

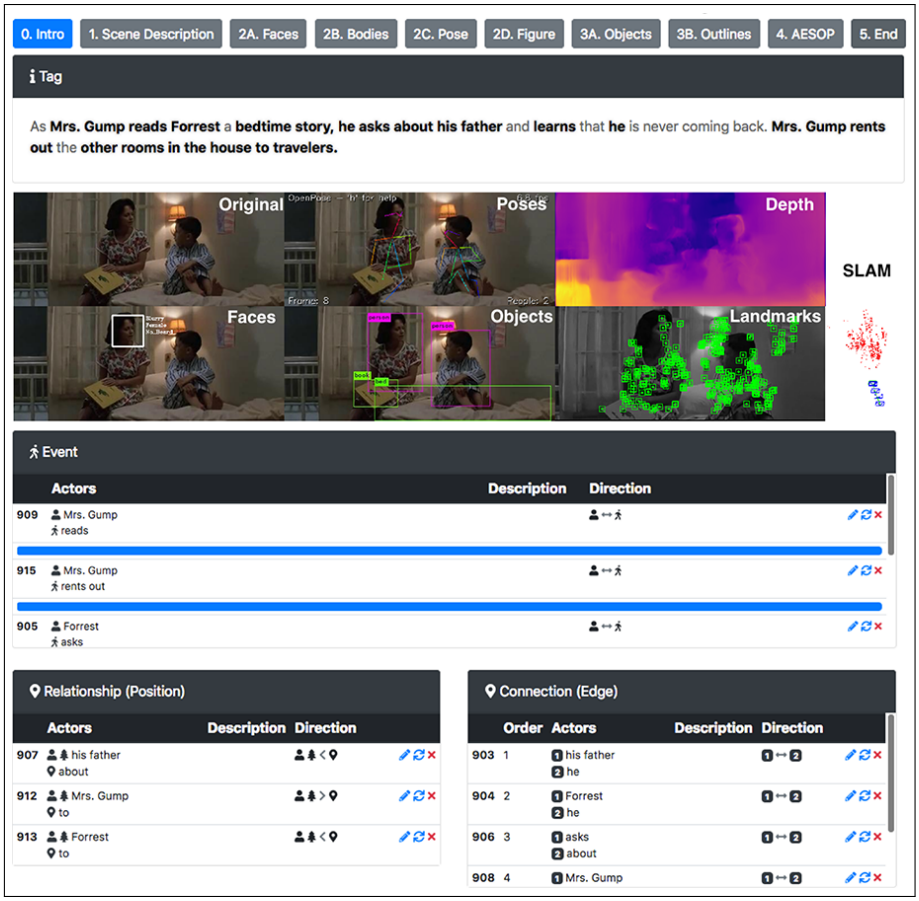

Figure 4.1: Aesop's annotation mode with commonsense grounding component, consisting of a number of interface components supported by a language parser and various computer vision pre-processing tools. The user, in sequence, can focus on different aspects of the film scene to establish and augment knowledge graphs, which can be later used to ground the AI system and recreate the scene via another animation software application. The user can additionally establish relationships between characters and objects, and indicate the sequence of events that take place in each scene.

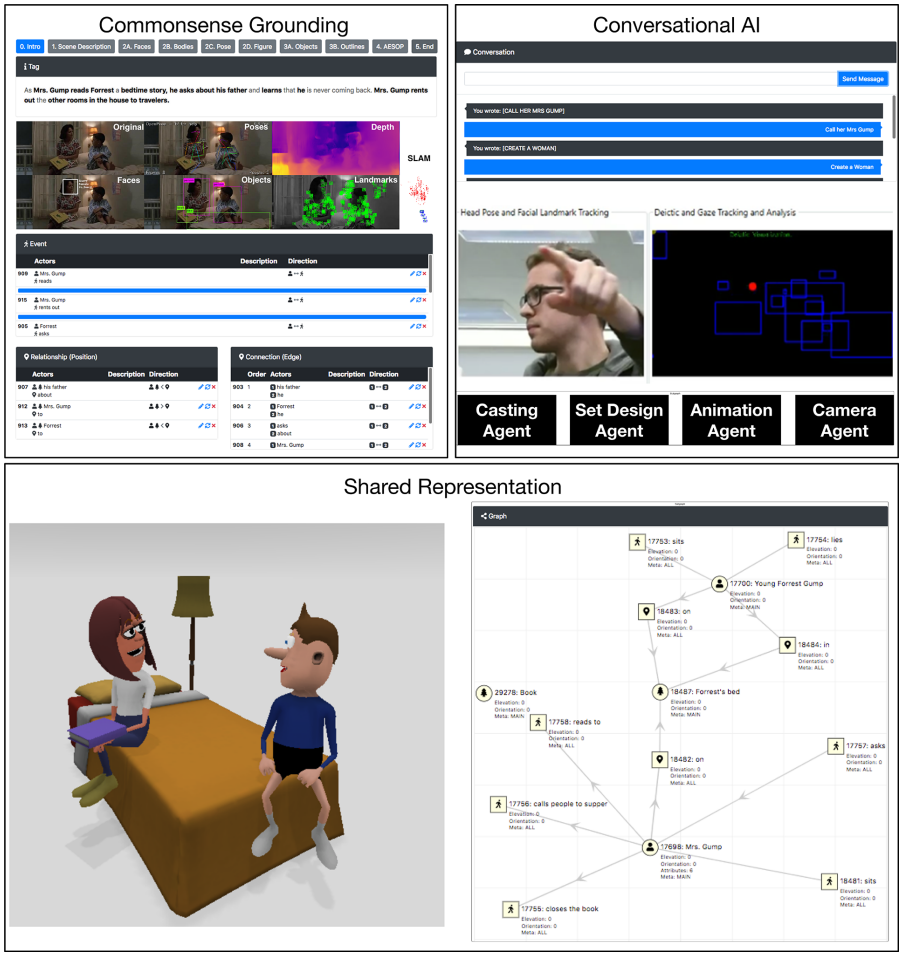

Figure 4.2: An brief illustration of the overall Aesop system, an end-to-end visual storytelling platform that allows the user to deconstruct individual film scenes using an annotation system and instruct the AI system to recreate the film using different interaction modalities. The top section illustrates Aesop's two main modes — annotation with commonsense grounding component and collaboration with conversational AI mode (top right, further expanded in a separate chapter) — and the bottom illustrates the visualized animation (bottom left) and the underlying knowledge representation graph (bottom right). All aspects of the system was built with Modular's default and project-specific modules, with Muvizu providing visual storytelling assistance.

4.3 System Architecture

Built using the previously introduced Modular framework, Aesop features two main modes as illustrated in the top section of Figure 4.2: commonsense grounding responsible for collecting annotation from human users, and conversational AI that relies on grounded concepts from the former.

Both modes are based on the underlying spatio-temporal knowledge graph, which serves as an intermediate representation between language, vision and animation. This graph, represented as a complex network of nodes and edges, is communicated to an external animation software Muvizu [114] as a series of API calls, visualizing the system's understanding of a scene as a pre-rendered, full-motion animation.

4.3.1 Annotation with Commonsense Grounding

In order to capture human insights surrounding film scenes, script segments, and individual on-screen entities and synthesize otherwise discrete AI detector outputs, Aesop provides a flexible interface as part of the commonsense grounding mode. Displayed in Figure 4.1, Aesop's commonsense grounding mode consists of a number of underlying components: a language parser [136] that accepts a user-provided story to build a semantic representation; various computer vision pre-processing tools, which parse a video and produce object, face, and pose detection, along with depth and tracked landmark points for simultaneous localization and camera mapping. The user can use available detector output in order to establish and augment knowledge graphs, which can be later used to ground the AI system and recreate the scene via the licensed animation software, Muvizu. Recognizing the interaction cost involved in deconstructing and annotating different aspects of each film scene, the annotation interface provides page-based navigation that allows the user to focus on specific data types or contexts, and each page features its own set of relevant modules.

4.3.2 Collaboration with Conversational AI

Built to take advantage of accumulated commonsense to construct new and original film scenes, Aesop's conversational AI mode, shown in Figure 4.2 (right), consists of an equally complex set of components: a speech language parser [4], which takes speech input and constructs a semantic graph; an external gesture tracking application that interprets depth and color information of the human user's camera image to produce the corresponding deictic coordinates; a dialog manager, which parses user inputs, tracks the trajectory of individual parsing processes, and adds them to the shared knowledge graph. This mode engages various movie-making agents responsible for analyzing the knowledge graphs and triggering corresponding API calls to Muvizu. The components relating to the conversational AI mode and its collaborative capabilities will be further discussed in the subsequent chapter.

4.3.3 Animation Software

To further visualize its knowledge graph in the form of a conventional animation movie, Aesop relies on a separate software and its preset props and animations called Muvizu [114]. Designed to emulate an actual movie production pipeline, Muvizu cuts out many of the barriers associated with the creation of an animated film and assists Aesop for visual storytelling. Muvizu provides a large library of built-in assets to rapidly assemble a scene, and enables users to select characters, customize their appearance, position cameras and lights around the set. In turn, the users can quickly prepare for each scene's key shots, issue directions to the actors, and record on-screen action in real time.

Its internal library of pre-generated assets includes 80 characters, 600 props (complete with the ability to import custom 3D object file), 1000 character accessories, 6 types of lights, 900 pre-animated character actions with mood-based modifiers ("pointing angrily" vs. "with fright"), and 19 visual effects for cameras. Shots are layered with visual and audio effects, voice tracks, and music, and finally exported as a video file. A user can direct character eye and head movement, and automatically lip sync characters with audio tracks. All internal assets are directly exposed via API endpoints, enabling its interaction

Figure 4.3: Muvizu software mimicking an instance from a movie.

with external applications such as Aesop. Figure 4.3 illustrates Muvizu's ability to replicate a frame from a movie to a relatively realistic level of detail.

4.3.4 Knowledge Graphs

Defined as a directed graph representation of a scene, Aesop's knowledge graph is used to encapsulate the spatio-temporal and object-attribute relationships within a scene — and serves as an intermediate representation between natural language, vision and the animation domain. Inspired by Modular's graph module, actors and props are represented as discrete modules, each associated with various descriptors and attributes. As illustrated in Figure 4.4, these nodes are bound together with links that correspond to spatial relationships, and temporal relationships such as interactions with other actors and actions.

Aesop uses the knowledge graph as the core representation of the corresponding scene, and implements a bidirectional interface between each module in Figure 4.4 to the central knowledge graph. Users can also incrementally build the graph using natural language in the conversational AI mode, as Aesop can extract relationships from the output of the speech language parser to add, remove, or modify nodes in the graph. Finally, users can also ground the knowledge graphs on Muvizu and visual and textual representations from movies by highlighting visual or textual tokens creating a knowledge graph that is directly linked to Muvizu assets.

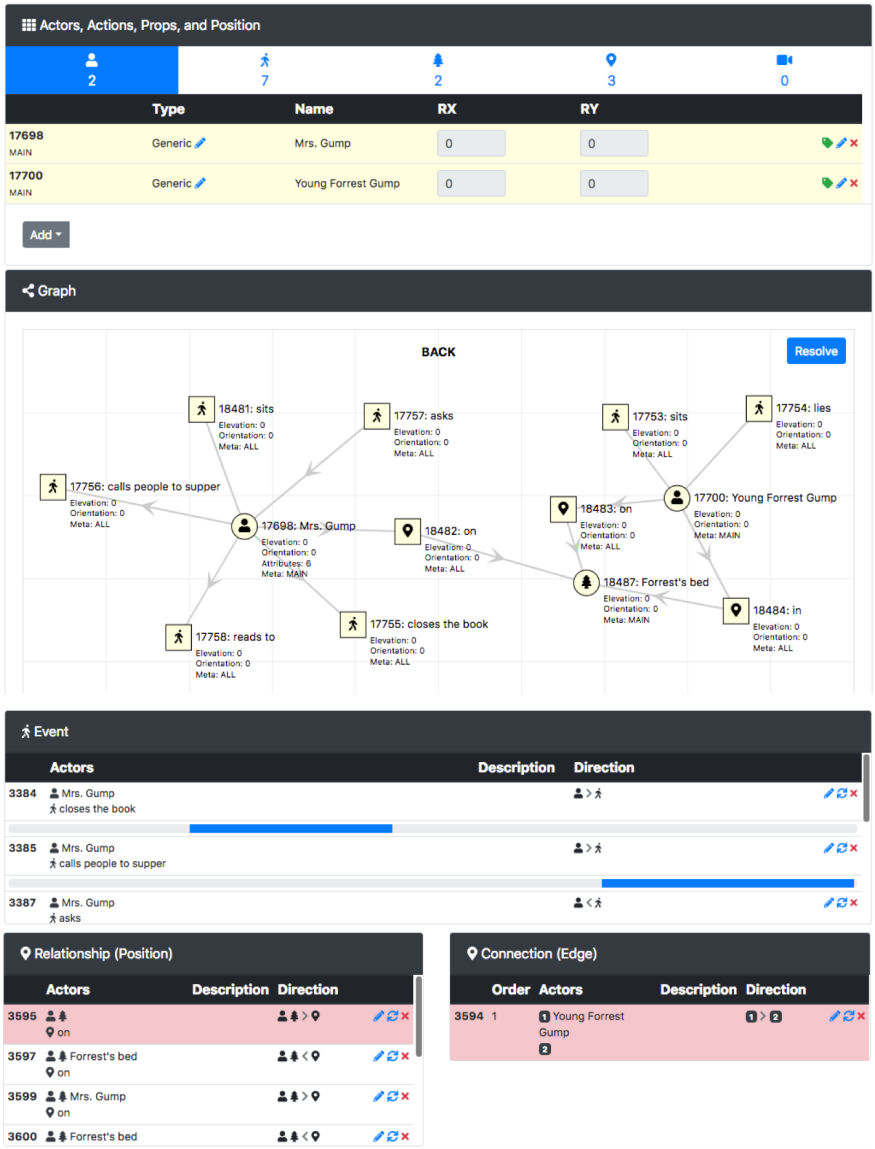

Figure 4.4: An example of a knowledge graph with the various entities (people and props) and relationships (spatial and temporal). The "Actors, Actions, Props, and Position" module (first) displays all actors, their various actions, and other relevant scene components; The "Event" module (second) indicates which specific actions were performed by individual actors, complete with start and end time; the "Relationship" module (fourth) defines positional links between actors and props, as well as interpersonal links between individual actors; finally, the "Graph" module (second) presents a summary view of all nodes and links. Though contextually rigid in this interface, all entities share the same data schema in the backend and are amenable to new entity types as required. The current set of available entity types was informed by Muvizu software capabilities.

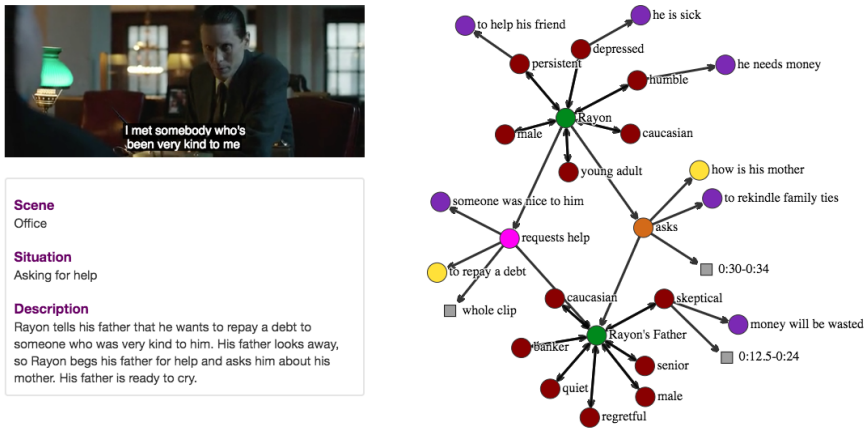

Figure 4.5: This figure is a courtesy of MovieGraphs dataset [155]. Each instance comes with a video, textual description of the scene, a situation label and a graph representation of the situation.

4.4 Commonsense Grounding

Building upon prior work in commonsense grounding and visualization of narratives and scripts, Aesop's annotation mode utilizes a number of components to collect human insights and establish a complex knowledge graph of film narratives: previously available datasets, pre-processed textual and visual assets, and a web-based annotation platform.

4.4.1 MovieGraphs Dataset

The MovieGraphs dataset [155] is a collection of 7,637 annotated video clips from 50 movies. Movies are a rich source for human interactions: in the dataset, clips are annotated with characters who appear in the scene, complete with their various attributes (both physical and emotional) and relationships and interactions between each other. The dataset also features timestamped actions performed by individual characters, as well as a brief description of the scene's narrative content.

Illustrated in Figure 4.5, each situation instance includes a video, subtitles, a brief description of the scene, and a corresponding situation graph. Currently, MovieGraphs are annotated with human-centric annotations, lacking

objects, spatial relationships and common sense grounding. Aesop relies on the MovieGraphs dataset as an initial source of knowledge and commonsense, and allows the system's users to further augment it with spatial relationships, visual features extracted from the source video, and user-provided annotations — all to teach the AI agents how to translate visual and textual concepts into grounded knowledge graphs.

4.4.2 Textual Data Processing

Aesop augments the MovieGraphs dataset using information extracted from text data, such as entities, spatio-temporal relationships, and context, and utilizes SLING [136], a frame-semantic parser for natural language, for story text parsing. SLING uses a semantic frame that represents a collection of facts, features, and attributes of a detected object and its relationship to others. The resultant data is again processed and used to further train the parser.

SLING was an advantageous choice for Aesop as it can flexibly adopt different schema for representing entities and their relationships in different ways. For instance, one schema that the pre-trained model understands is the PropBank schema [77]: a corpus of text annotated with information about basic semantic properties. PropBank annotations, when adapted by SLING, enable the parser to identify both the subject and the object affected by the subject. The benefit of using the PropBank schema is using the links between frames to identify the corresponding subject and object or relational arguments of the event in question. Furthermore, one can use the graph output of SLING and PropBank to generate text based knowledge graphs for reasoning over spatio-temporal object-attribute relationships. Once an event is extracted, corresponding nodes and edges are automatically created in the shared graph representation illustrated in Figure 4.4.

4.4.3 Visual Data Processing

Aesop also uses additional visual features extracted from the MovieGraphs dataset. In addition to the available graphs, each clip was pre-processed by

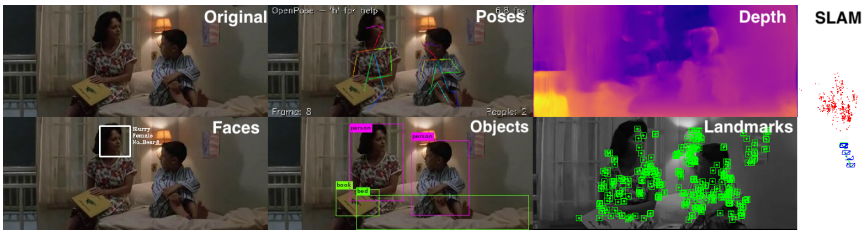

Figure 4.6: Preprocessing of a video clip from MovieGraphs. Each video is processed through: a face detector, tracker, and attribute classifier; human pose and facial landmarks estimator; a monocular depth estimator; and a landmark detector for SLAM (simultaneous localization and mapping) of the camera.

running it through an additional set of detectors. Illustrated in Figure 4.6, a number of unique detectors run on individual frames from the available clips.

YOLO [132] is a real-time detector designed to detect humans and inanimate objects, and provides context, enabling the users to understand more about the surroundings of the actors and better ground the scene layouts (the movie set). Facenet [5] and MXNet [23] are used to detect human faces, identify individual characters and their facial attributes, such as gender, facial hair and hair color. This detector enables extraction of character attributes and grounding them in Aesop. Openpose [20], a human pose and facial landmark detector, is used for activity detection, as this detector detects human skeletons, where activity nodes are grounded on. Finally, Simultaneous Localization and Mapping [113] is used to track the camera's path throughout the scene's 3D space. This is achieved by detecting persistent landmark points and tracking them. This detector enables understanding of the camera relationship with objects and actors, as well as its motion through the scene.

Each of the detectors creates a corresponding node in the knowledge graph, augmenting the current MovieGraphs dataset with additional context. As movies are a two-dimensional projection of a 3D world, Aesop requires depth estimation of the scene. Monocular depth estimation [51] is used to detect the relative distance from the camera to objects and surfaces in each frame of the shot, and this information is made available to all other detectors to facilitate the commonsense grounding process. Individual outputs from the above AI detec-

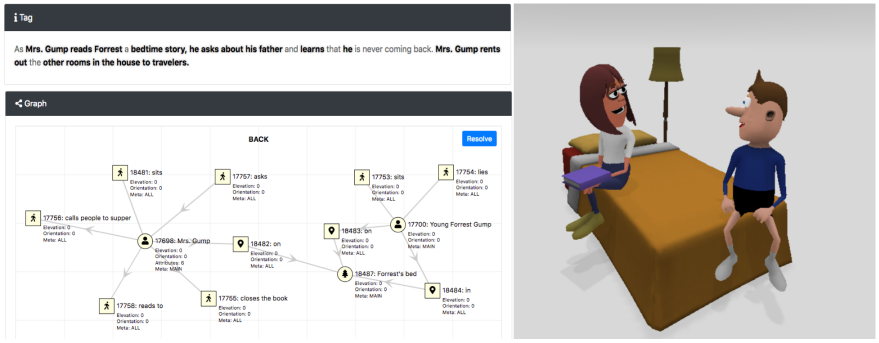

Figure 4.7: An example of grounding a scene in Aesop. The sentence is about "Mrs. Gump reads Forrest a bedtime story." This would entail that Mrs. Gump and Forrest are in a bed, most likely sitting or laying on bed, with Mrs. Gump holding a book and reading from it. In grounding such knowledge graphs, created from textual and visual knowledge, augments the knowledge of the AI clarifying the implicit information.

tors have been manually parsed, transformed, and collated for the use on Aesop per standard Extract-Transform-Load (ETL) procedures common in business intelligence activities [152].

4.4.4 Grounding on Aesop

As multiple datasets converge into Aesop, the user is invited to join the commonsense grounding process and contribute to human-in-the-loop automation. The augmented MovieGraph dataset is then visualized alongside the source video clip using the web-based interface, illustrated in the right top section of Figure 4.2, where users can manually map those graph nodes to entities within Aesop. The mappings in the grounded graph, illustrated in Figure 4.7, are saved for future retrieval — allowing the AI system to learn from them and continue to resolve nodes as required.

Character nodes map directly to Aesop's Character-type (e.g. "Young Forrest Gump" maps to "Boy"). Each of the nodes gets its casting information from the detected character attributes and mapped to the corresponding Aesop entity's attributes e.g., hair color, shirt color, height, etc., provided they exist within Aesop's asset library. Object nodes are instantiated as Aesop's prop type (furniture, vegetation, decorations, etc.). Monocular depth values are averaged

across an object's bounding box to determine the order of proximity to the camera relative to other objects in the scene.

Action and interaction nodes, grounded on skeletal data from OpenPose along with their timestamps grounded via Aesop's animation engine, are matched to the list of available actions within the Aesop animation library. These decisions are based on the action node description, and the selected character actions are executed at the designated timestamps within the scene.

Finally, camera tracking from the ORBSLAM detector provides a motion path for Aesop's camera node, over which Muvizu provides full control. The grounding of MovieGraph nodes via entities in Aesop's 3D world can take one of three forms:

- **One-to-one**: A single MovieGraph node maps to a single Muvizu entity, e.g. "Young Forrest" maps to Aesop's "Boy" character.
- **One-to-many**: A single MovieGraph node maps to multiple Muvizu entities, e.g. "Principal's Office" maps to "Desk," "Chair," "Lamp," etc.
- **Many-to-many**: Multiple specific MovieGraph nodes are understood to map to multiple Aesop entities, e.g. "Forrest walks over to the doctor" resolving to two character nodes and an action node.

While annotation conflicts are rare, the interface prevents potential erroneous mapping (e.g. "Young Forrest" mapping to "Lamp") using a dialog window that presents the preexisting mappings and further confirms user actions.

4.4.5 Web-Based Annotator

The web-based interface additionally provides a suite of tools that enable the users to refine previous annotations and create new annotations as necessary. Building upon the original iteration [104] designed with expert users and a command-line interface in mind, this Modular-based interface allows users to perform such tasks with minimal training using a familiar web browser.

Consisting of simple word token tagging and free-form text input components, text modules enable each user to mark different terms or create comments

pertaining to a specific part of a text-based dataset. These user-provided annotations can also introduce an element of collaboration as the interface offers a visual indication of previously tagged terms, as well as a visualization of agreement amongst the uploaded entries. The tagging module tokenizes individual sentences and paragraphs to enable token-based tagging, while the free-form module allows the user to submit free-text comments.

Video annotation modules offer a variety of methods to annotate individual frames in a single video clip: bounding boxes and polygons can be created to mark different on-screen entities, while landmark points can be used to represent skeletal and facial feature data. These annotations are presented as overlays above the original frame, and allow the users to directly move and resize. In addition, the visualization below the video player displays a series of dots across the timeline, indicating occurrence and co-occurrence of individual entities.

Finally, the graph module allows for visualization of knowledge graphs that underpin individual scenes using node-link representations that encapsulate spatio-temporal and object-attribute relationships. Just as importantly, the module also allows for directly updating links between such nodes using a mouse cursor. This module is especially instrumental in grounding discrete datasets such as MovieGraphs by allowing human users to intuitively map nodes between discrete datasets.

4.5 Application

An external collaborator with deep experience in script writing and video editing was recruited to demonstrate the utility of Aesop's capabilities as an annotation platform. Over the course of three and half months, the domain expert was able to accurately ground the previously disconnected datasets — MovieGraphs, SLING and PropBank, and various image detector outputs — across all 162 scenes of *Forrest Gump*. With the exception of initial tutorials and adhoc technical troubleshooting due to occasional user error and software bugs, the expert was able independently complete the data annotation and commonsense grounding process without relying on other annotation software. A partial result of this

annotation process is illustrated in Figure 4.4, where the annotator was able to identify the on-screen characters, observe their actions, and establish positional and interpersonal links between the characters. Accumulated annotation data was then used to demonstrate the collaborative creative process available in the conversational AI component of Aesop, further elaborated in the subsequent chapter.

4.6 Summary

Recognizing prior work and present opportunities in establishing generalizable knowledge base for artificial intelligence, Aesop, a visual storytelling system, introduces a novel end-to-end commonsense grounding process. This chapter can be summarized as the following:

- Previous efforts in commonsense grounding lie in simplistic and rigid environments bearing little resemblance to the physical world.
- Aesop's knowledge graph encodes scene information with flexibility for augmentation and modification by the human user, and fuels the AI agent's deeper understanding of the scene.
- Aesop's annotation mode, as evident in experience with an external collaborator, establishes a case in augmenting previous datasets with external detectors and a web-based interface.

Chapter 5

Human-AI Collaboration in Content Creation Tasks

Artificial intelligence is becoming good at many "human" jobs — diagnosing disease, translating languages, providing customer service — and it's improving fast. This is raising reasonable fears that AI will ultimately replace human workers throughout the economy. But that's not the inevitable, or even most likely, outcome. Never before have digital tools been so responsive to us, nor we to our tools. While AI will radically alter how work gets done and who does it, the technology's larger impact will be in complementing and augmenting human capabilities, not replacing them.

H. James Wilson and Paul R. Daugherty, *Collaborative Intelligence* [163]

5.1 Introduction

Beyond the debate of AI-driven automation and technological unemployment, there exist complex tasks where human and AI parties can work together to complete: AI agents may be responsible for reviewing a large set of available documentations and making the necessary recommendations, as the human user focuses on the creative endeavor of completing a work of visual design; conversely, the human user may submit a series of speech commands to the AI worker, which then proceeds with building the product.

While interaction between humans and machines commonly focuses on handling directives given by the human user, the future of artificial intelligence holds potential for more mixed-initiative collaboration where human users and AI agents act as equals [33]. An AI agent may try to communicate a goal to a human collaborator, evaluate the decisions made by the human user, and intervene or contribute additional ideas as an equal in a process.

The two previous chapters respectively focused on gathering human insights for commonsense grounding and improving explainability of AI decision-making processes and system inner-workings. In a next natural step, this chapter sets out to explore the domain of symbiosis between human users and AI assistants by focusing on two case studies in application of human-AI collaboration to content generation tasks. Aesop's content generation mode presents an opportunity for the user to take the helm of a film director with AI assistance, and MUSICA provides an application to interact with a musically talented AI system to generate music and even perform alongside it in real time.

5.2 3D Animation Content Generation

As discussed in the previous chapter, Aesop primarily serves the goal of facilitating collation of discrete AI detector outputs and user-provided annotation and resolution of knowledge graph. On the other hand, *Aesop* can alternatively function as a system with the goal of content creation by conversing with a set of AI agents using verbal and non-verbal communication to co-create animations. Aesop provides a rich platform that enables research in language, gestures, vision, and planning in the context of storytelling. Aesop uses shared knowledge graph representations created from language and vision, using the Modular-powered interface, to generate a 3D animation sequence. The user also can engage with Aesop and receive corresponding animations using a chat window similar to today's messenger applications. This section elaborates on the conversational AI mode of Aesop, consisting of the speech language parser, the gesture and gaze analysis module, and the conversational module.

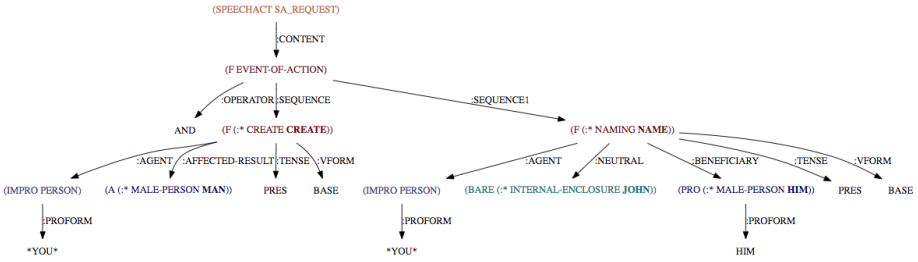

Figure 5.1: TRIPS parser logical form of "create a man and name him John."

5.2.1 Speech Language Parser

Aesop uses a symbolic language parser TRIPS [4], a broad-coverage domain-general deep semantic parser that produces a Logical Form (LF) grounded in a general ontology. It generates a semantic representation structured around events. TRIPS uses a general language-level ontology, augmented with domain-specific knowledge about the visual domain. The output LF is a directed acyclic graph of ontology tokens representing entities, events, and their spatial, temporal, and lexico-semantic relationships. The LF is used to determine a grounded problem-solving act that represents a common goal between the human and computer. We illustrate a speech parse in Figure 5.1 showing an example of a LF for a speech parse: "create a man and name him John." The LF is an encoding of the semantic content of a sentence that can be mapped to a knowledge representation. TRIPS also identifies the agent performing the act and the objected affected by the act.

5.2.2 Gestures and Gaze

Similar to prior work on incorporation of pointing gestures and gaze for linking semantic entities with objects [74], communication with physically embodied agents [100], and disambiguation of expression reference and reference resolution using non-verbal communication [146], Aesop offers its own component that addresses such challenges.

Aesop integrates a Multimodal Integrated Behavior Analytics (MIBA) system illustrated in Figure 5.2. MIBA allows Aesop to watch for the user's gestures

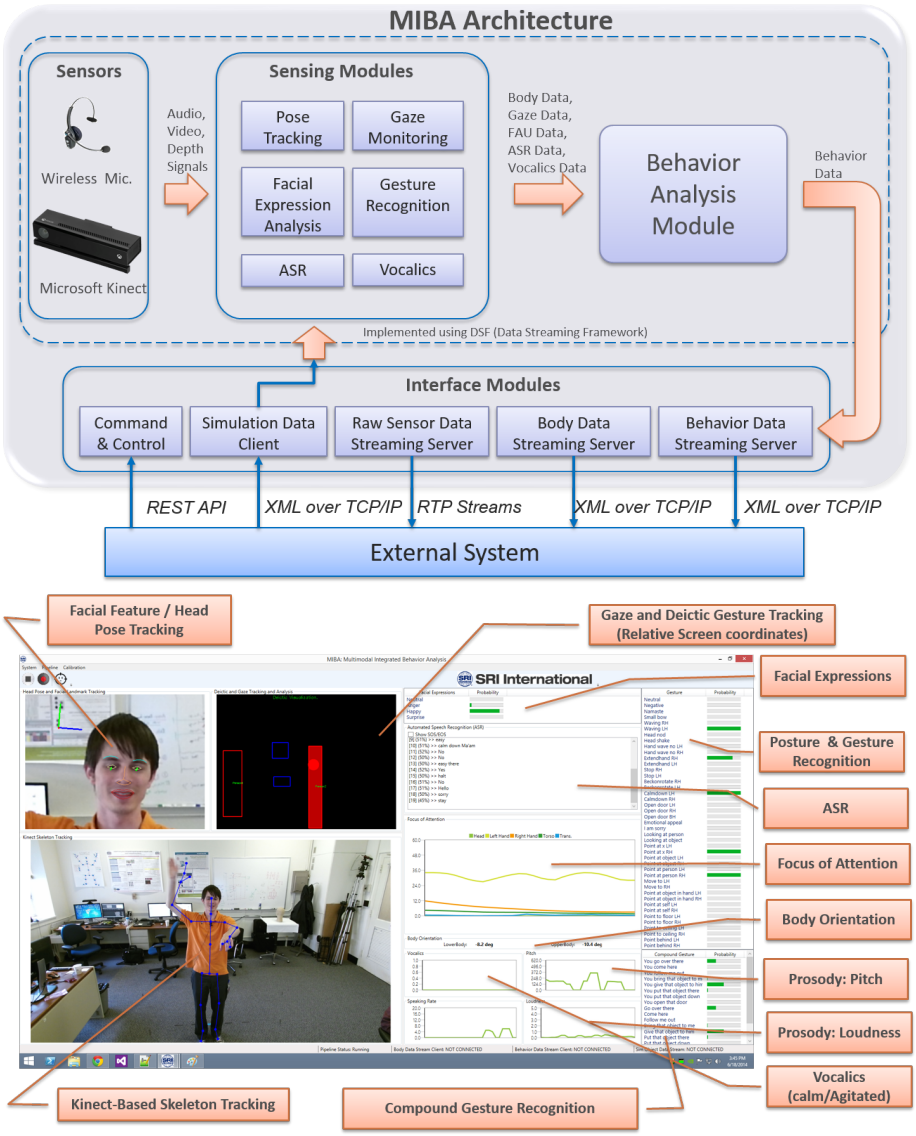

Figure 5.2: Multimodal Integrated Behavior Analysis (MIBA) architecture (top) and a dashboard of MIBA analytics showing the different non-verbal behavior analytics performed (bottom), developed by another project collaborator. Though not designed as the primary method of interaction with the Modular-powered Aesop, the user can use this alternative method to interact with Aesop and manipulate objects in Muvizu animation software.

and gaze in concert with the verbal utterances to manipulate the objects and characters in Muvizu, and to do so, MIBA relies on a Microsoft Kinect as the underlying sensor. The sensor provides a high resolution RGB video stream, a 3D depth video stream of the person, and a high quality audio stream via an on-board microphone. For noisier venues, we can utilize a wireless lapel microphone for the audio stream in place of the Kinect audio.

The user interacts with Aesop by sitting or standing in front of a large screen. Emblematic gestures and the 3D deictic gestures arising from the user's pointing at things and user's gaze on screen emerged as the most utilized MIBA components during the prototyping process. MIBA integrates with the object stream from the Muvizu software in the form of bounding boxes on the screen. The 3D deictic vectors from the deixis and gaze are projected onto the Aesop display and their intersections with the object bounding boxes allows MIBA to identify the objects of reference and objects of attention. This makes it possible to resolve references to specific objects and locations within the Muvizu world to place actors and props exactly where the user (director) wants them to be placed. The deixis also allows the director to explicitly define (i.e., draw) paths for object and actor trajectories within the Muvizu animations.

5.2.3 Module for Conversational AI

Using the web-based controller, the user can engage with Aesop and receive corresponding animations based on its interpretation of the user's request. Built on popular web technologies and hosted on a publicly accessible server, this platform-agnostic interface allows the user to interact with Aesop without the need to use a dedicated client application.

Featuring a familiar chat window similar to today's messenger applications, the interface allows the user to submit a text input and take advantage of the text parsing mechanism constructed by a project collaborator. User-provided text is then committed to a database, which triggers Aesop to utilize TRIPS to parse the text: as per the previous example, "create a man and name him John" is tokenized, flagged, and converted to an internal graph representative of the scene. Upon completion, Aesop responds with a series of messages describing its

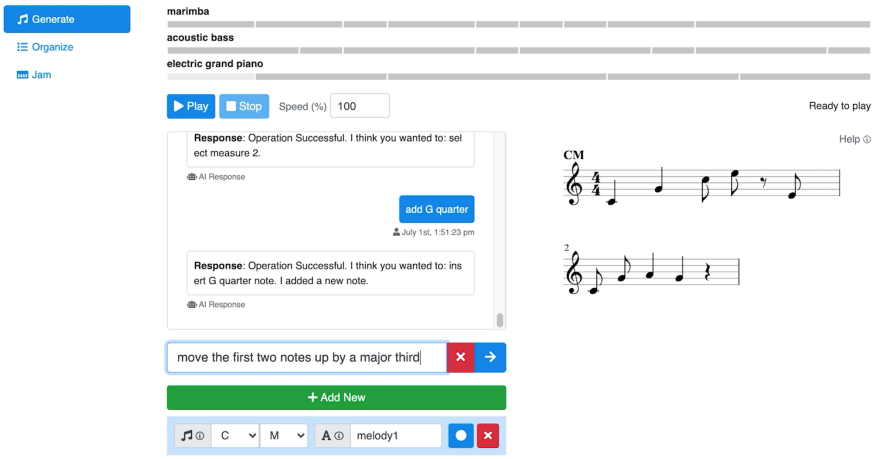

Figure 5.3: The "Generate" tab of MUSICA, where users can create short melodies using a chat-style interface or a traditional point-and-click interface.

parsing process and prompts the user to review the results: a knowledge graph displaying various nodes and edges representative of the scene as described by the user, and Aesop's reconstruction of the very same scene using Muvizu.

The user can then choose to continue interacting with Aesop using the chat window, or directly interact with the resultant nodes and edges (corresponding to Aesop's internal state) to further refine the graph. The interface provides a number of modules that allow the users to create manually create nodes and edges that are compatible with Muvizu assets, and Aesop updates its animation according to the user's additional updates.

5.3 Interactive Jazz Generation

MUSICA (MUSical Interative Collaborative Agent) is a project that focuses on interaction and communication between human musicians and machine assistants. Built on computational models of music and natural language processing for musical operations [126, 129], the project offers different methods in which the human user can interact with machine to generate new music, as well as various opportunities to inform future research regarding musical language and human-computer interaction.

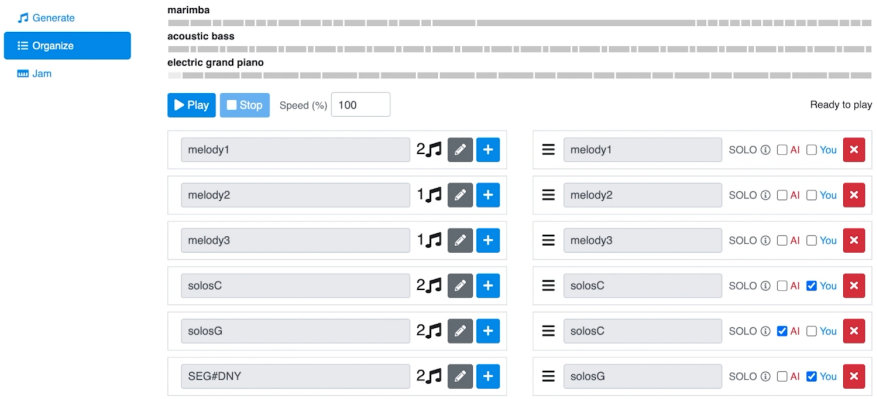

Figure 5.4: The "Organize" tab of MUSICA, where users can collate, manipulate, or delete existing musical segments to build a longer piece of music, visualized with varying-length gray blocks representative of individual musical notes (top). Users can freely construct the composition using available segments with drag-and-drop (bottom left), and toggle between between AI and human performance modes (bottom right).

The user can ask the algorithm to automatically generate bars of music, remove a specific set of musical notes, and listen to the work-in-progress with simple text commands using a familiar chat window [125]. The user can interact with the same piece of music using other methods, including using a mouse pointer to manipulate individual notes or playing a specific piece of music with a MIDI controller. Finally, the user can organize a collection of completed musical segments into a larger piece and practice using the interface, alone or with AI accompaniment that respond to user's performance patterns. The web-based interface, once again, is powered by the Modular software platform.

5.3.1 Music Composition

Unlike most existing music software, MUSICA features a natural language interface to create scores. Users start by creating one or more short musical segments, where each segment can have a melody and an accompanying chord type. Illustrated in Figure 5.3, the primary tool for creating and editing each segment is a chat-style interface. Those with limited musical experience can leverage the system's generative capabilities and high-level transformations to easily create music. MUSICA also features a help system to offer suggestions

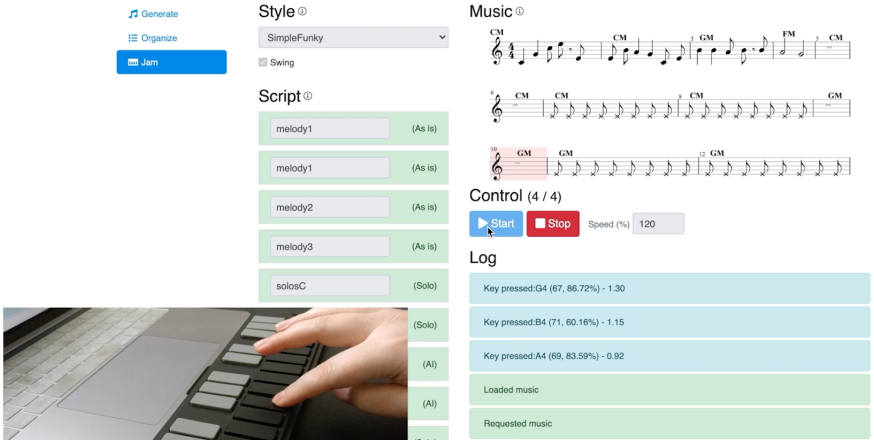

Figure 5.5: The "Jam" tab of MUSICA, where users can improvise with the system in real time using the keyboard or an external MIDI controller. The system monitors the user's performance in real time, and uses the collected notes and the corresponding segment's musical key as seeds for latency-free generative music.

when it doesn't understand the request and to familiarize users with the various operations available to them.

Built on MusECI (elementary composable ideas), a specialized framework designed to establish a language for score-level music representation [128], the chat-style interface relies on the aforementioned TRIPS parser to process natural language commands, select a specific part of the music, and manipulate accordingly. More advanced musicians can assume a greater level of control by using domain-specific terminology to add or change specific notes in the chat. Users have the ability to manipulate the score at more than one level of granularity, working on whole measures, individual notes, or collections of notes. The system also accommodates a range of syntactic variation to support different preference for how to phrase commands: the user may ask the system to "add a C note," "raise its pitch by one octave," and "move the note to the second bar" in three separate queries, or simply request the system to "add a C3 note to the second bar." Finally, a point-and-click interface offers an alternate way to accomplish some of these changes.

5.3.2 Real-time Improvisation

Once a collection of musical segments has been composed, the user can organize them into a larger piece and then perform it as illustrated in Figure 5.4. Users can leave the performance entirely up to the computer, or they can improvise with the system in real time and trade solos in several different playing styles.

Extending prior work in grammar-based generative music [127] based on user-provided input [67], the AI system stands by as the improvisation session begins and the human user starts to perform using the device of choice. As the performance approaches the boundary between the end of the human turn and the start of the AI counterpart, individual MIDI messages are passed to the AI agent, which quickly generates the musically compatible set of notes represented as the corresponding MIDI messages. These messages are queued for synchronized playback, and the result is a real-time exchange of jazz solos.

Illustrated in Figure 5.5, the user has access to an interface that indicates the current state of the improvisation experience using the segment list, the log window, and the classical music notation. The user may change the tempo or the accompany style prior to commencing the jam session, and individual measures and notes change their colours in synchronization with the rhythm of the music in order to visualize the pace and the progress of improvisation. The performance can also be generated entirely by the computer, based on the user's preferred style and each segment's musical key.

5.4 Summary

Inspired by a range of use cases in human-AI collaboration for mixed-initiative tasks, this chapter presents two distinct incarnations of the Modular framework and demonstrates the following:

- 3D animation generation with AI assistance is possible with the use of natural language parsing, human gesture integration, and chat interface.
- Musicians can compose and improvise music with AI by combining representation framework, natural language processing, and music generation.

Chapter 6

Approach to Improved AI Explanation

Explanations are used to manage social interactions. By creating a shared meaning of something, the explainer influences the actions, emotions and beliefs of the recipient of the explanation. For a machine to interact with us, it may need to shape our emotions and beliefs. Machines have to "persuade" us, so that they can achieve their intended goal. I would not fully accept my robot vacuum cleaner if it did not explain its behavior to some degree. The vacuum cleaner creates a shared meaning of, for example, an "accident" (like getting stuck on the bathroom carpet ... again) by explaining that it got stuck instead of simply stopping to work without comment.

Christoph Molnar, *Interpretable Machine Learning* [112]

6.1 Introduction

As consumers increasingly entrust their personal information and security over to AI solutions, lawmakers continue to push back on industry players and call for more regulation [28]. On the other hand, AI models continue to absorb human-provided data to further automation, and concerned commentators recently talk of existential risks imposed by such technologies. While these debates surrounding trust and reliance on artificial intelligence continue, a more

fundamental question, troublingly, often goes unanswered: did we ever fully understand the decisions made by artificial intelligence over the years?

Transparency and explainability seem have emerged as important attributes in artificial intelligence, as we focus on how data-hungry and powerful AI decision making processes have become. While AI has become ubiquitous in consumer products as they make a number of decisions for us, ranging from film recommendations to advertising preferences, it is no secret that many of us continue to treat an AI system as a black box: the system's algorithms suffer from lack of transparency, as it is difficult to determine the internal mechanism functions other than to infer based on a collection of inputs and outputs. In life-altering, high-stake decisions such as disease diagnosis or legal determination, it is crucial that these predictions and results must not simply resemble those made by human counterparts, but also provide reasons behind such decisions: after all, it is difficult to trust a system that cannot explain itself [1].

Explainable Artificial Intelligence (XAI) has emerged as an area of potential solution in response to interest in AI systems and their ethical conundrums. XAI enables new machine learning techniques, specifically deep learning, to yield explainable models. These explanations can be developer-focused (to help in understanding, designing, and improving models) or user-centric (to help in knowing how and when to trust the outputs of AI tools). From the user-centric point of view, it is of crucial importance to explain the decisions of an AI system with effective explanation techniques to enable end users to understand, appropriately trust, and effectively manage the AI-originated decisions. An effective XAI system assists in the human decision-making supported by the system, in particular whether to accept the recommendations or classifications suggested by the model.

In modern AI systems, the most critical and most opaque components are based on machine learning. There is an inherent tension between machine learning performance (predictive accuracy) and explainability; often the highest performing methods, such as deep learning, are the least explainable, and the most explainable, such as decision trees, are the least accurate. From a decision-making point of view, the goal of XAI systems is to maintain performance while

being explainable.

The target of XAI is an end user who depends on decisions, recommendations, or actions produced by an AI, and therefore needs to understand the rationale for the system's decisions. For example, a test operator of a newly developed autonomous system will need to understand why the system makes its decisions so that they can decide how to use it in the future. A successful XAI system should provide end users with an explanation of individual decisions, enable users to understand the system's overall strengths and weaknesses, convey an understanding of how the system will behave in the future, and in some cases even suggest how to correct the system's mistakes.

Inspired by practical challenges of visual search and ranking in areas of commerce and surveillance, this chapter explores the paradigm of "explanation by generation" using a novel generative XAI system for human activity search and ranking in motion capture data. Recent work on visual search and ranking largely focuses on black-box discriminative methods [7], where the system searches for a specific video clip given a query with little to no insights into its mechanism. The presented XAI system, on the other hand, features a more transparent mechanism based on the Dense Validation Generative Adversarial Networks (DVGANs) approach [92]: given a query, the system generates multiple video hypotheses and use them to search for the answer. As a result, the system provides the user an insight on what the model "thinks" the query looks like using a web-based interface — instilling an element of explainability. This web interface, as was the case for Aesop, was built upon the Modular software platform.

6.2 Related Work

The XAI system, consisting of an explainable model presented using a user-focused dashboard interface, is inspired by prior work in three main areas of research: the recently emerged XAI domain, generative adversarial networks, and information visualizations conducive to interpretability and explainability.

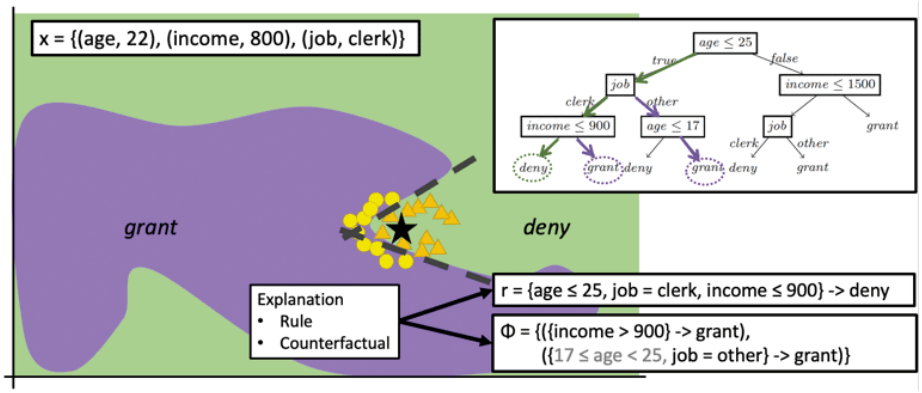

Figure 6.1: Sample explanation of a black box AI classifier [55]. The visualization provides clear division between two possible outcomes and explanation of which inputs are responsible for such outcomes via different types of visualization. While this visualization may be useful in illustrating the decision-making process of the AI system, it may not be readily interpretable or conducive to understanding by non-expert audiences.

6.2.1 Explainable Artificial Intelligence

Recent work produced in the XAI domain as well as new opportunities that emerged therein serve as the main underpinnings of the generative system. Traditionally, as various inference systems extended their capabilities, there has been a need to trace and represent each system's decision-making process in order to justify its conclusions, identify any contradictions, and further improve the corresponding operations [30]. In response, there has been a demand for more transparent, explainable AI systems as an extension on the need to understand both automatically and manually coded rules.

Global explanations focus on analyzing overall learned representations: for example, understanding and visualizing representations in deep learning (e.g. convolutional neural networks) [73, 118, 169], analyzing representations learned by deep reinforcement learning agents (e.g. deep Q-networks) [168] or learning disentangled representations [63]. In the global explanation case, after the model is learned, the explanation is then extracted from the representation learned by the model itself, as illustrated in Figure 6.1.

Local explanations focus more on grounding the explanations on specific data, for example, finding influential features [134] and grounding them on the

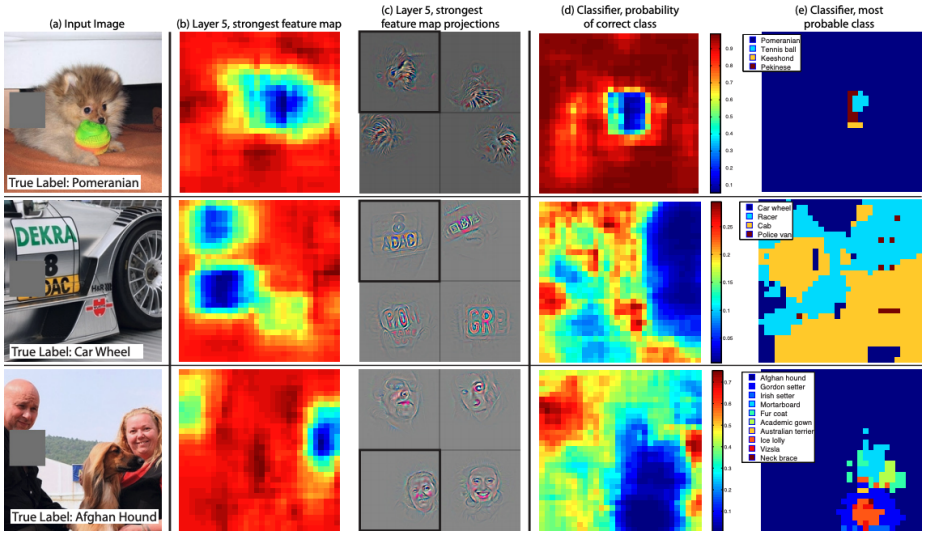

Figure 6.2: Sample visualization of convolutional neural networks in relation to the input image and the corresponding results [169]. Influential features are identifiable in the form of heatmaps and highlighted edges, though not explicitly announced as such for non-experts.

input image as illustrated in Figure 6.2. Other methods focused on finding influential data points [79] and parameterizing training batches, and some focused on generating textual explanations by training a second deep network to generate explanations without explicitly identifying the semantic features of the original network [59]. Finally, attention-based methods for explanation, such as show and tell networks, couple captioning with attention on images [157], using attributes for attention [22] or using guided attention [87].

Beyond the above developer-centric, expert-oriented explanations that focus on AI models and corresponding data points, however, there is also a call for more intuitive and interpretable explanations for human users. Some XAI projects pursue "human-in-the-loop," user-centric systems that produce trustworthy answers without significantly compromising the system performance [81, 134]; other applications seek different ways to ensure fairness and accountability by providing users alternative outcomes using counterfactual statements ("had a number of conditions been different, the outcome would change") using intuitive visual interfaces [53] and voice assistants [149]; finally, ideal systems also

provide contextually relevant recommendations and explanations to their end users who may have little to no technical knowledge in AI, but are experts in their own domains [57].

6.2.2 Generative Adversarial Networks

Generative Adversarial Networks [54], which the model of this chapter's generative XAI system is based on, are a class of implicit generative models that learn directly from examples. Employed successfully in many problems, GANs are active mostly in the area of computer vision where they are trained directly on pixels. There are multiple variations of GANs, many of which propose a variation of the objective function to address different needs. Starting from the original formulation [54], the extension to Conditional GANs (CGAN) [49] was introduced to enable conditioning on a class label; Wasserstein GANs (WGAN) [6] was introduced to improve the stability of GANs; finally, WGANs with Gradient Penalty (WGAN-GP) [56] improved WGAN's stability even further by replacing weight clipping with a gradient penalty in the loss function.

The system specifically draws from approaches for human motion generation, which features two main types of synthesis: (1) motion completion, starting from a short clip and extrapolating to a longer clip, and (2) motion generation, starting with a label and generating full clips. Recent work on human motion modeling for motion completion successfully used recurrent neural network (RNNs) [48, 50, 69, 99], although human motion generation relied on previously available dataset instead of from scratch. Recently, GAN-based approaches have been applied successfully to synthesize human motion from text [2, 9, 92] by formulating a sequence-to-sequence model using a GAN framework [99].

6.2.3 Explanation Interfaces

The system's explainable interface is largely inspired by recent advances in interpretability, trust, and explainability in the information visualization and human-computer interaction fields [24, 25, 93]. Information visualizations are often populated with the outputs of machine learning techniques, however, simply visualizing the outputs of an ML system is insufficient as an explanation.

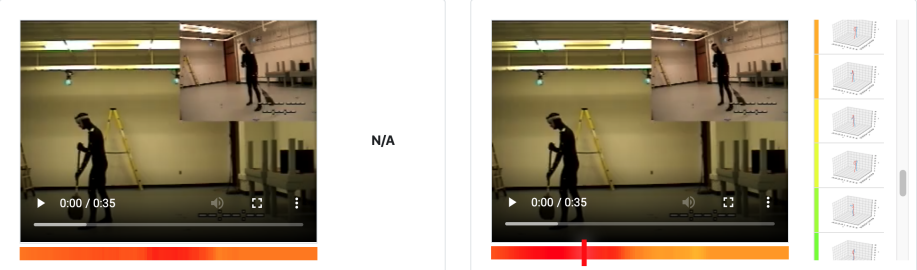

Figure 6.3: Comparison of two video search engines powered by two ranking algorithms: discriminative ranking (left) and generative ranking (right). The discriminative model (left) is trained to rank a list of videos by extracting features from them and relate to the text query. On the other hand, the explainable model (right) is trained to generate video instances of what it thinks the text query should look like, and use these instances, accessible in the right sidebar, to rank the videos. Each vertical bar in the list encodes spatial information pertaining to the sweeping action in each clip, specific to the selected point in video timeline: red-orange hues indicate that the action is most likely present in that particular point in time, while green-blue hues indicate the opposite.

Where visualizations such as a topic model plot [40], rendering of features in convolutional neural networks (CNNs) [93], or a t-SNE (t-distributed stochastic neighbor embedding) model visualization [154] may be useful for those familiar with the workings of the algorithm, they are inappropriate for users of the system with little expertise in machine learning as indicated in preliminary expert case studies [40].

Provenance is a key consideration for supporting decision-making in data analytics, and providing traces of both data provenance and analytic provenance has been used to enhance the trustworthiness of analytic outcomes using visual analytics [139, 162]. Analytic provenance tools have recently been the focus of much visual analytics research, and are often a variation of an automatically populated storyboard showing the history of interaction [43, 171].

6.3 Visual Search and Ranking

Recognizing the opportunities in “opening the black-box” with explainability, the XAI system, dubbed GenRank, tackles the challenge of visual search and

ranking with generative models inspired by GANs, pitted against the black-box, discriminative AI model as briefly illustrated in Figure 6.3.

6.3.1 Challenge

Human activity understanding is a rich area of research in robotics, computer vision and machine learning, due to the challenges it offers. GenRank, however, focuses primarily the surveillance use case [7].

Search and ranking, a process that involves querying a video database for certain activities of interest, is a data-intensive exercise: due to the large size and number of video frames, instead of searching the pixels directly, each video is processed by extracting visual abstractions such as objects, parts and their spatial configurations. This is usually achieved using detectors [20, 132, 151], but by using a database of captured motions where human body joints are accurately localized in 3D using motion capture devices, one can conveniently access the spatial motion of human body parts connected to the query of interest.

CMU Mocap [153] is a large-scale motion capture dataset of open-ended activities. It contains 2,548 high frame rate motion capture videos from 113 actors performing 1,095 unique activities, complete with text descriptions. There are activities with different styles and transitions such as "walk on uneven terrain," "dance - expressive arms, pirouette," "punch and kick," and "run to sneak": Having such fine-grained activities brings us close to human activity understanding in the wild, and CMU Mocap is the largest dataset of its kind.

Available in the BVH format [58], the dataset features human body skeletons represented by 31 joints, closely resembling those from the previously introduced DVGANs approach [92]. The joint angles are pre-processed into the exponential map representation and activities spanning less than 8 seconds are filtered out. The filtered dataset contains 573 actions across 1,125 videos totaling 8 hours. We use 757 videos for training the AI retrieval system and 368 videos for evaluation.

The AI retrieval systems spot query activity using frame-by-frame sliding windows of 8-second clips. As part of this work, two state-of-the-art deep models are compared: (1) a discriminative ranking model which does not provide global

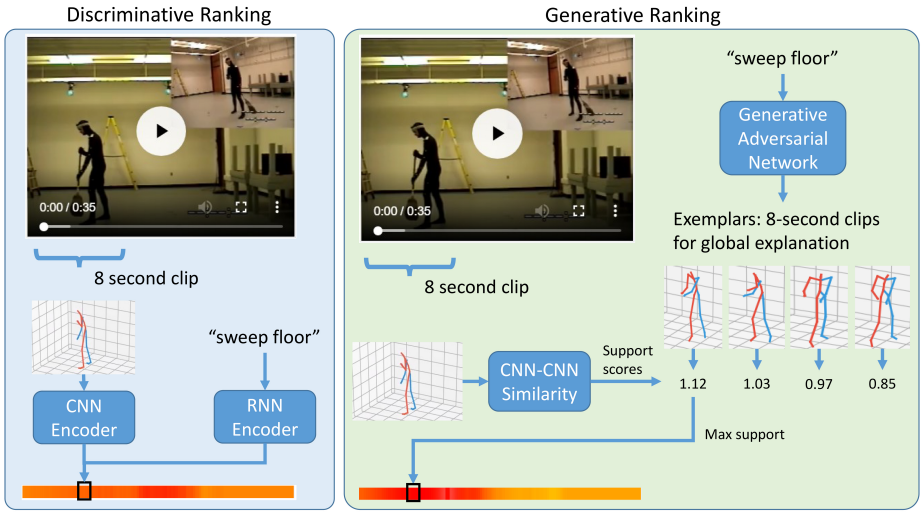

Figure 6.4: Illustration of discriminative ranking vs. generative ranking. (left) Given a fixed-length video clip in the video database and a query "sweep floor," discriminative ranking uses a CNN-RNN model to score the clip. (right) Generative ranking first generates exemplar clips of what the model thinks is "sweep floor" then uses the clips to score the video database through a CNN-CNN similarity function. The score is indicated by the confidence bar, where red indicates a higher level of confidence.

explanation, and (2) a generative ranking model which provides exemplar-based global explanations about the retrieval decisions. The discriminative ranking and generative ranking systems are illustrated in Figure 6.4.

The discriminative ranking algorithm may provide a competent solution to this problem, and this may be perfectly acceptable should the answer happen to be correct. It is unclear, however, exactly how the model may have arrived at a certain solution, should the solution be incorrect, leaving the user confused. Can the generative ranking algorithm perhaps explain to the user its inner-workings?

6.3.2 Discriminative Ranking Implementation

Inspired by the state-of-the-art model for ranking image captions [44], the discriminative ranking model, illustrated in Figure 6.5, computes a ranking score of how well the clip matches the query. The input action text is encoded into a query vector using the skip-thought vectors model [78], and fed to a GRU

(gated recurrent units) RNN language model. The input video clip is encoded into a video clip vector using a 1D residual convolutional neural network, and the matching score between the query and video clip is computed as the dot product between the query vector and the video clip vector. The AI system, powered by this model, selects videos with highest average score over all its aforementioned 8-second sliding window clips as the output. The ranking scores for sliding window clips are visualized to the user to explain the retrieval decisions — and model parameters are learned such that matching score is high when text matches the video, and low when the text does not.

The discriminative ranking model has been trained, with the help from another project collaborator, jointly for action classification and retrieval of human activity videos. The action classification task is: "given a fixed-length video clip, retrieve its original description from a pool of 250 descriptions." Similarly, the action retrieval task is: "given an action description, retrieve the video clip that corresponds to the action description from a pool of 250 fixed-length video clips." The negative log-likelihood action retrieval and action classification losses are optimized to learn parameters of the CNN, via the Adam optimizer with learning rate of 1×10^{-4} over 100 epochs.

6.3.3 Generative Adversarial Networks

In a typical GAN, there are two components which learn as adversaries: a generator and an discriminator. The generator is tasked with creating videos which will fool the discriminator into classifying them as real videos of the target action, while the goal of the discriminator is to assign high values to real videos and low values to generated videos as a "video appraiser."

Training of this system is through a cyclic game in which the discriminator learns to improve its classification performance, followed by the generator learning to improve the quality of the fake video it creates. In traditional GANs, the end output is a high accuracy discriminator which can detect and classify videos representing an action. In the approach used for the XAI system, the outputs of the generator are also examined as representations of how the system understands the supplied action term.

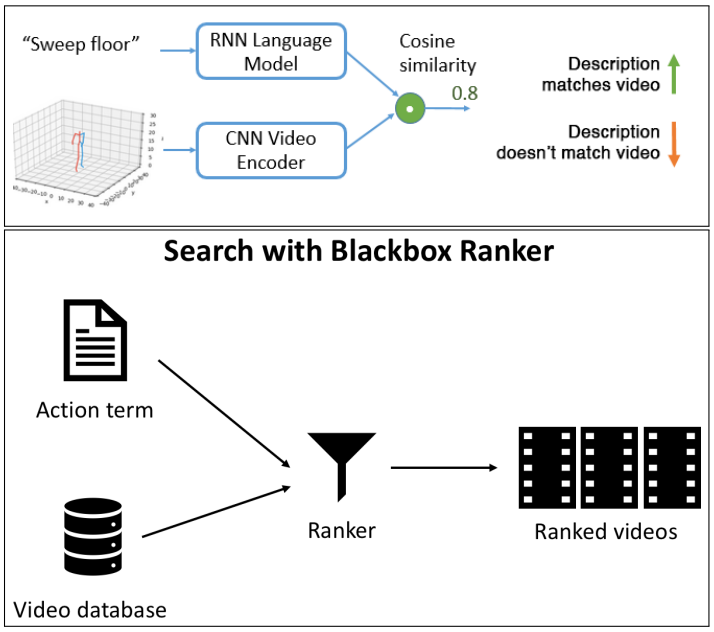

Figure 6.5: Illustration of the black box AI system designed as a counterpoint to the generative XAI system. Given a query and a video clip, discriminative ranking computes a ranking score, computed with cosine similarity, of how well the clip matches the query (top). Based on this mechanism, the system can parse through the video database and create a list of ranked videos.

The objective function is the average score assigned to real videos of the target action, minus the average score for generated videos. The discriminator is trying to maximize this objective, while the generator is trying to generate more realistic videos and minimize this objective.

The extension to Conditional GANs (CGAN) for face generation [49] concludes that when the game reaches equilibrium, the generator will generate the real video distribution and the discriminator will not be able to tell fake videos from real videos. The generative model is learned using Wasserstein GAN [6], which greatly stabilizes GAN optimization.

6.3.4 Generative Ranking Implementation

Given a query, the generative ranking model first generates exemplar clips using a text-conditioned GAN [92]. For example, for a query "walking," the model will generate a set of exemplar clips which it "thinks" represent the "walking" action. In successful cases, when our generator generates videos close to the distribution, sufficient to "fool" the discriminator, the generated videos will capture all kinds of different pose, style and timing variations. The user can examine the outputs of the generator to understand how the system interpreted the concept, as illustrated in Figure 6.6.

Since the users have direct access to visualization of the synthetic examples, the only black-box component left is the similarity function between a synthetic example and the input video. The system uses Euclidean similarity, which we believe is sufficiently intuitive for the users to understand. Similar to discriminative ranking, videos with the highest average score over all sliding window clips are retrieved as the output. This particular method is ideal for the user-facing interface that allows the users to query by natural language, as this is the only GAN-based approach that requires a text query to generate a video clip without example frames as an additional input. For generative ranking, the generated exemplar clips are global explanations of the retrieval decisions. The generated exemplar clips are presented to the user along with the ranking scores to explain the retrieval decisions. To clarify, exemplar clips are visual representations of the dynamically constructed body joint coordinates based on the generative model. While we have an option to use an existing 3D model to map to these coordinates for immersion and visual fidelity, we chose to illustrate these data points as minimalist "stick figures" in a three-dimensional grid to minimize potential distraction and visual obstruction.

6.3.5 Performance

Both models are benchmarked by their top-1 accuracy when ranking 248 8-second video clips, equipped with unique action descriptions, in the CMU Mocap test set. With random ranking of $1/248 \approx 0.4\%$, the discriminative ranking approach achieves 35% top-1 accuracy, while the generative ranking approach

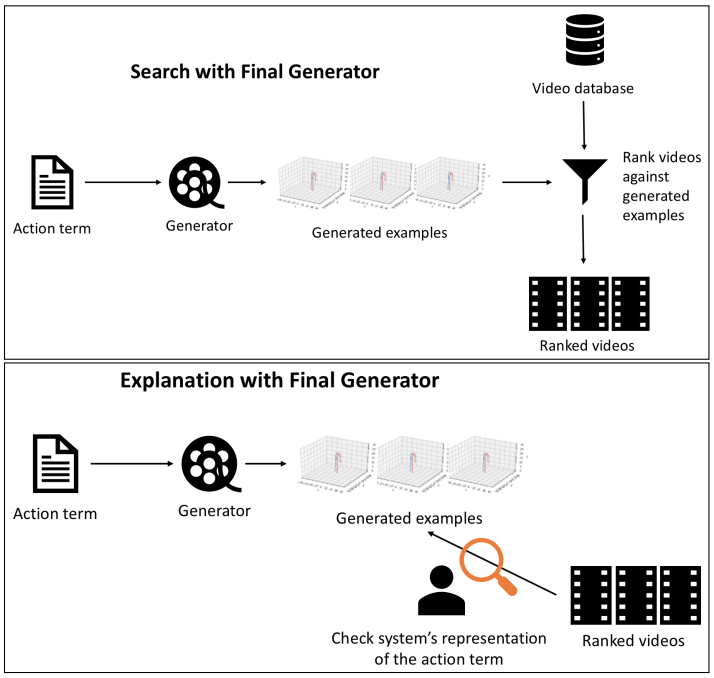

Figure 6.6: Illustration of the generative XAI system in action. Given a query, the system generates many synthetic examples that correspond to the query through calling the video generator mechanism (top). The system then uses each synthetic example as a query to further look for similar videos in the dataset and aggregate the top matches (bottom). Intuitively, this step is trying to find the highest similarity from any synthetic example to input video. The scores are then normalized to the range of [0-1] for visualization.

achieves 33% top-1 accuracy. Therefore their performance is comparable. For global explanation, discriminative ranking provides only the matching scores, while generative ranking, in addition to scores, naturally generates exemplar-clips for the query that can be shown to the user. Finally, generating 5, 10, and 20 exemplars resulted in top-1 retrieval performance to drop by 1.6%, 1%, and 0.5% respectively compared to 30 exemplars.

6.4 Explanation Interface

The interface acts as a mediator between the human and the AI, to help understand the AI's rationale for decisions through a variety of explanation ap-

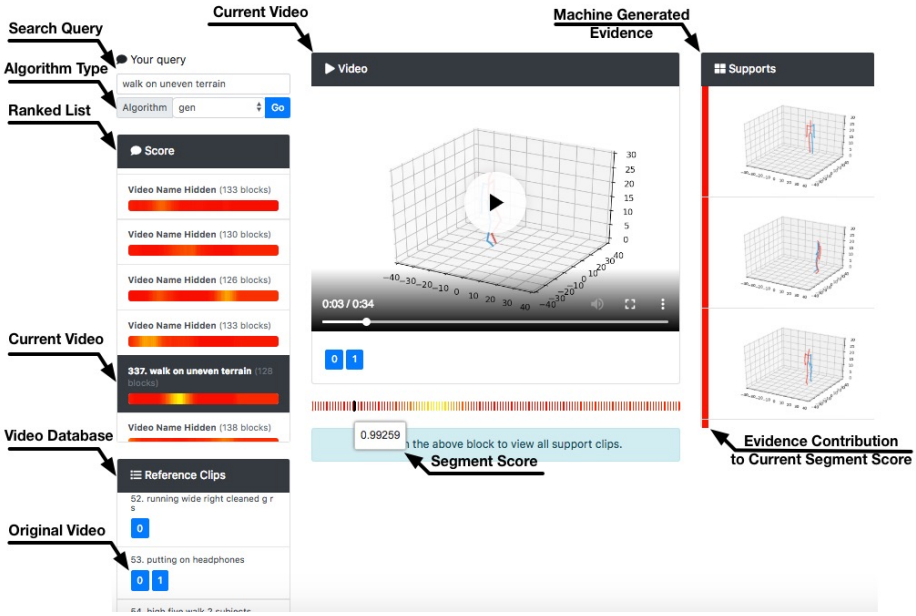

Figure 6.7: Annotated view of the explanation interface. In the left column, the user starts by inputting a query and selecting the ranking algorithm of interest. In the middle column, once one of the videos is selected from the ranked list, the video becomes available for the user to view, along with a segmented confidence bar (red indicating higher confidence) located below the navigation bar. In the right column, only in the generative ranking case, the user is also presented with an unsorted set of generated evidence used to rank the list of videos. Upon clicking on the confidence bar in a certain segment, the score for that segment is presented to the user, and the list of evidence is sorted showing the most important evidence first. A pop-up video player is presented to the user to view each video.

proaches. The explanation interface combines visualization of generator instances from the AI generator, as well as uncertainty in the outcomes from the ranker. Looking to set out to bridge the gap and make the explanations appropriate for people who are not AI experts, the system features familiar constructs such as animated motion sequences, rather than abstract visualizations of hidden model states or other low level features. The dashboard adopts a surveillance use case where the target user is an analyst searching for an activity in a large video database.

The explainable interface assumes a dashboard design geared towards domain experts who may not have deep knowledge of artificial intelligence, but

are well-versed in the data upon which the system operates. Purpose-built to visualize AI rationale for individual video clips in the dataset, the interface is also designed to handle similar application areas and accompanying datasets, enabling users to answer textual queries with visual rationales and confidence scores for the answers. Illustrated in Figure 6.7, the interface also allows a data domain expert to observe the evidence and make an informed decision whether to trust the XAI system — by supporting "drill-down" into deeper evidence for the provided answers and the level of confidence the model is reporting.

The interface enables the user to input a search query, choose one of the matching keywords from the training set, and select a ranking algorithm (*Generative* or *Discriminative*). Once the search button is clicked, all the videos in the database are ranked in a new list with the most relevant video on top, and each video is equipped with a *confidence bar* visualization. Once one of the videos is selected, the video becomes available for the user to view, along with the same confidence bar visualization located below the navigation bar, showing areas of interest where the query term is most likely to occur. In the case of generative ranking, in addition to the confidence bar, the user is also presented with an unsorted set of *generated evidence* that the algorithm used to rank the list of videos. Upon clicking on the confidence bar in a certain segment, the score for that segment is presented to the user, and the list of evidence is sorted to show the most important evidence first. Upon clicking on one of the evidence clips, a pop-up video player is presented to the user to review the generated evidence in detail.

The generative ranking interface enables the user to drill down into deeper evidence for the provided answers and see the level of confidence the model is reporting. This allows a data domain expert to make an informed decision whether to trust the system. In the discriminative ranking case, the user would have little evidence to support the provided decision. The ability to drill down on evidence ends at the ranked list and confidence bar.

With this interface, a question quickly emerges: how will an explanation interface lead a user to accept or reject the answer provided by the AI model? This research question is discussed in the subsequent chapter of this disserta-

tion, as the user study investigates the role of trust and reliance, and whether generated evidence engenders appropriate trust in AI systems.

6.5 Summary

Designed to improve explainability of AI output without compromising performance in comparison to the black-box counterpart, the generative XAI system sets out to provide AI assistance with visual explanations and a familiar web-based interface. This chapter can be summarized as the following:

- Transparency and explainability have emerged as important concepts as humans increasingly rely on AI.
- Conventional AI models are considered black-box and their mechanisms are not conducive to human interpretation.
- Generative XAI system generates a series of video queries based on the user request, which serve as explanation elements in its behavior.

Chapter 7

Assessment of Trust and Reliance in Human-AI Collaboration

Part of the reason we trust other people as much as we do is because we by and large think they will reach the same conclusions as we will, given the same evidence. If we want to trust our machines, we need to expect the same from them. If we are on a camping trip and both simultaneously discover that the eight-foot-tall hairy ape known as Sasquatch (aka Bigfoot) is real and that he looks hungry, I expect you to conclude with me, from what you know about primates and appetite, that such a large ape is potentially dangerous, and that we should immediately begin planning a potential escape.

Gary Marcus and Ernest Davis,
Rebooting AI: Building Artificial Intelligence We Can Trust

7.1 Introduction

With AI systems being deployed to more high-stakes domains such as healthcare, criminal justice, and military operation, human users are expected interpret AI inferences and make informed decisions [8]. As the potential damage of incorrect

AI inferences becoming more significant, the "black box" nature of conventional AI systems continues to be challenged to provide further clarity over its inner workings and rationale for its answers.

Despite the allure of a more explainable, transparent system resulting in a stronger bond between human users and AI assistants, the question of whether such AI systems truly will help the users to understand the AI's mental model and benefit from improved performance of tasks at hand remains: after all, regardless of the system's ability to describe its mechanism, the user may prefer the conventional, black box AI system that can provide the necessary response without slowing down to explain itself.

Designed to evaluate the benefits of using the XAI system, this chapter presents the evaluation of the generative XAI system in the context of surveillance, where the user is querying a database of videos, searching for a specific activity. We use the CMU motion capture database [153], and devise an extensive user study that evaluates the quality of explanations, satisfaction with the explanations, and assesses the users' mental model, trust and reliance on the system.

7.2 Related Work

Designed to closely evaluate each participant's interaction with explanations, our user study is informed by a wide range of work surrounding user trust and reliance in AI explanations, as well as data-driven recommendation systems.

Example-based explanations are generally considered an acceptable way to rationalize algorithmic behaviour [18, 80], while participant reactions to these explanations vary greatly and are subject to individual differences, including self-confidence and prior experience with explanations [35, 142]. Users also generally prefer to retain control over the application, leaving recommendation and AI-driven behaviour only accessible by request [38]. Users are also most likely to opt into utilizing AI assistance in order to quickly identify the answer, with little interest to learn how or why AI arrived at its solution [37, 120].

When users do accept automatic assistance, however, they do so at arm's

length: distancing their own decisions from algorithmic behaviour and adjusting trust levels according to the accuracy of systems [166]. Finally, past work has observed general indifference to how these explanations were presented, as there is no strong preference for the level of detail or the type of visualization presented in the explanations [37, 142].

7.3 User Study

The user study consists of various components to assess different factors from a mental model to trust and reliance. Designed as a more linear, guided variation of the explanation interface, the study presents numerous instances of three main tasks:

1. Identifying one or more video clips that best illustrate the displayed query.
2. Spotting one or more segments in a single video clip that best illustrate the displayed keyword.
3. Collaborating with AI to solve a more complex challenge of identifying a longer video clip that best illustrates a query with multiple actions.

The study sets out to assess the benefit, if any, of XAI by comparing two conditions: a condition where the participant uses the XAI system, and another where the participant takes on the same tasks using a black-box AI system. The two systems are powered by generative and discriminative models respectively, and while the two systems behave consistently, the XAI system prominently features the previously discussed model-generated video clips that illustrate the system's understanding of the keyword. The study also rigorously records the user's subjective experience with questionnaire components after each task.

7.3.1 Objectives

Our three-stage user study sets out to evaluate the model, the interface and the benefits of using the XAI system. Featuring a variety of interactive modules, this web-based interface was refined through an internal pilot and was deployed as part of a randomized controlled study, the results of which we report below.

Hypothesis. We hypothesize that the explanation interface will facilitate the user's understanding of the XAI system's behavior, while improving the user's task performance by building a correct mental model of the AI and establishing appropriate trust and reliance on the system.

Mental Model. A successful XAI system should allow users to gain a better understanding of the system's behavior, thus building a correct mental model of its operations. In this study, we use a series of prediction tasks and questionnaires to better understand the benefits of using an XAI system over a black-box one for mental model formation. Prior to gaining access to the XAI system's assistance, the user is presented with a sorted list of XAI-generated clips that illustrate how the system interprets the query. They are then asked to fill a short questionnaire to assess their expectations of the system. Given the presented clips, the user is asked to predict the decision of the AI on a specific task. The user's work is then compared to that of the XAI system to gauge whether the user was able to predict the system's behaviour. In addition to prompting the user with prediction tasks, the study also presents a number of assertions about the XAI system and asks the user to agree or disagree with the assertions. This way, their mental model is compared with an ideal model of the XAI system.

Task Performance. The user's task performance alongside the AI system is measured by comparing the resultant output with ground truth and observing the user's acceptance of the system. For each task, the participant has an option to view and use the AI system's output as the user's own. The study monitors the user's decision to accept the system's assistance, and examines the similarity between ground truth, the system output, and the user's answer. Collected measurements include task completion time, user and AI accuracy, and user reaction to the system-provided explanations as applicable per task.

Appropriate Trust and Reliance. Explanations should help users to develop more appropriate trust and reliance toward an XAI system and enable users to better achieve their goals. Maintaining close ties with the user's mental model of the AI system and resultant task performance, the study measures the

user's trust and reliance by asking the user to assess the level of confidence for the XAI system's output. The user can select an answer from a 5-level Likert scale, ranging from "Strongly Disagree" to "Strongly Agree," to indicate the user's confidence in the XAI system. The study also measures the user's reliance on the system by examining whether the user solicits the XAI assistance (through interaction log files) and continues to use it for subsequent tasks.

7.3.2 User Study Design

The study deconstructs the explanation interface into modular, guided user experiences to evaluate the benefits of using the XAI system over the traditional AI counterpart. Featuring numerous instances of three distinct tasks – *Clip Identify*, *Timeline Spot*, and *User-Machine Collaboration Task* – the study offers either the AI or XAI system to assist each participant along the way. Motivation and additional details surrounding the tasks are available in subsequent sections.

Overview. The study is a three-part experience featuring 2 different modes of AI assistance and 3 task types (3+3+2 repetitions) for a total of 8 AI-assisted tasks based on more than 40 different, randomly sampled configurations. The study is designed to take a maximum of 50 minutes to complete, and each prompted task is accompanied by Likert-scale questionnaires designed to record the user's subjective experience with the task. The study design and instructions were pilot tested with colleagues and students who were not part of the participant pool.

Participants. A total of 44 undergraduate students from computer science and information technology disciplines were recruited to participate in the between-groups study. The participants had no prior experience with XAI systems but had used commercial video search tools (e.g., YouTube). As there is no evidence that gender or age would be relevant factors, this information was not collected. Participants were compensated $20 for 1 hour of their time at the end of each session.

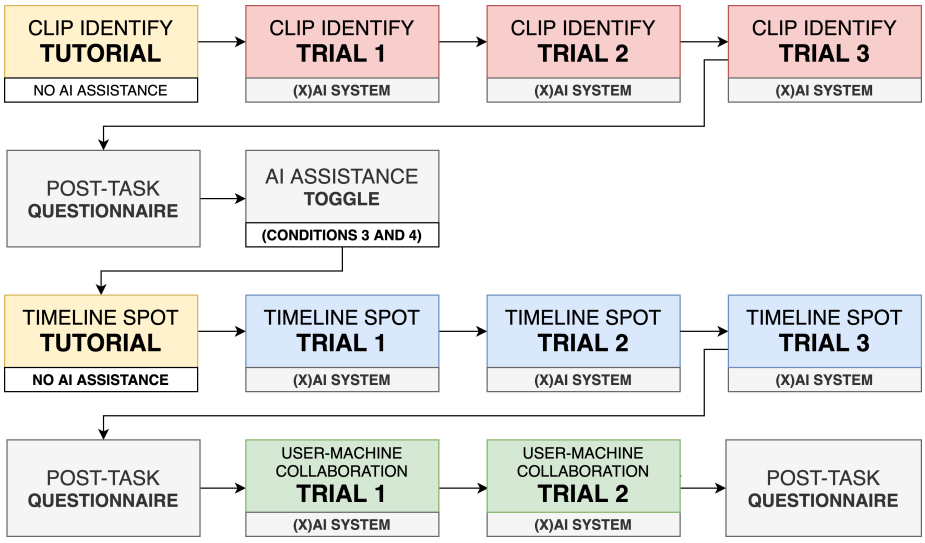

Figure 7.1: Illustration of the user study session flow. Throughout each session, the user performs numerous trials of three distinct tasks, each accompanied by a posthoc questionnaire specific to the task and the AI system's performance. Depending on the allocated experimental condition, the session may switch to the alternate AI system without warning.

Experimental Design. There were two conditions and three tasks. In order to alleviate order effects due to participant fatigue and practice with AI assistance, the following conditions were established:

1. AI system only
2. XAI system only
3. AI system, then switch to XAI system after "Clip Identify"
4. XAI system, then switch to AI system after "Clip Identify"

Each study session, dedicated to a single condition, was initiated with a brief introduction to the procedure and a tutorial about the study interface. 12 participants were invited to each session, with at least 10 participants successfully completing each condition, and no participants engaging in more than one condition. Participants worked individually on computers within a computing lab. The experimenter was available to answer participant questions throughout the session. Each participant received detailed training prior to beginning the study session, including completing the aforementioned sample tutorial tasks

and watching video recordings that illustrate ideal interaction scenarios.

General Structure. The study presents a number of text queries, accompanied by user tasks specific to each part, as well as applicable (X)AI assistance and questionnaire components. At the end of each task, the study also displays the summary that compares ground truth to user- and AI-provided answers. A detailed outline of the study is shown in Figure 7.1.

Part 1: Clip Identify. In this first part of the study, the participant is prompted to investigate a given set of ten video clips and pick up to three clips most relevant to the displayed keyword using a click-to-toggle interface as shown in Figure 7.2 (top). During this stage, the participant is presented a total of four trials with randomly selected keywords and associated identification tasks. The first trial serves as a tutorial and is not included in the results.

For each trial, the participant is first presented with the mental model questionnaire and asked to "step into the shoes" of the system and predict its answer. In the XAI condition, the mental model questions are accompanied by the sorted list of generated clips of what the system "thinks" the query looks like, as shown in Figure 7.3. After the mental model questions, the participant is presented with the task along with the results from assigned (X)AI assistance. The system's assistance is provided through sorting the list of clips, along with a confidence bar below each clip, showing the system's score over the length of the clip. The participant may view any clip during each task, and optionally import the system's suggestion as the solution. In the case of XAI, the user is also presented with AI-generated clips as evidence supporting the XAI system's interpretation of the text query. The evidence clips are rearranged automatically once a video is selected, according to which generated clips contributed the most to the system's decision. At the end of the trial, the participant is presented a summary showing the correct answer, their answer, their prediction of the system's answer based on their mental model, and the system's answer, as shown in Figure 7.2 (middle). Finally, the participant completes a questionnaire for this specific trial, assessing the performance of the system as shown in Figure 7.2 (bottom).

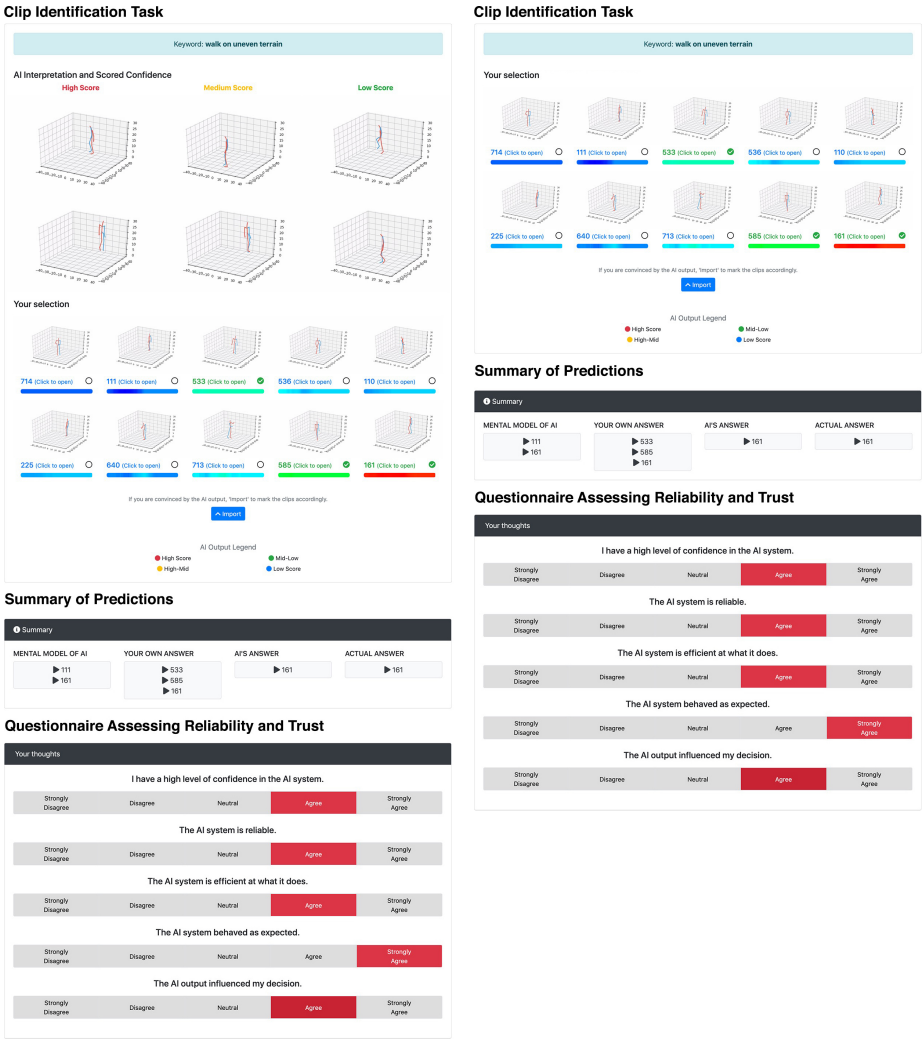

Figure 7.2: The *Clip Identify* task with the XAI system (left) and the AI system (right). Given a set of ten video clips, the user is asked to pick the top three most relevant clips to the displayed query. The XAI system, unlike the AI system, provides assistance with explanations using model-generated click-to-expand clips (top left). The options to select from appear below the the generated evidence (top left) or at the top (top right). Below that, a summary showing the correct answer, the user's answer, the user's prediction of the system's answer (mental model of AI), and the system's answer appears. (bottom) A questionnaire per trial assessing the performance of the system.

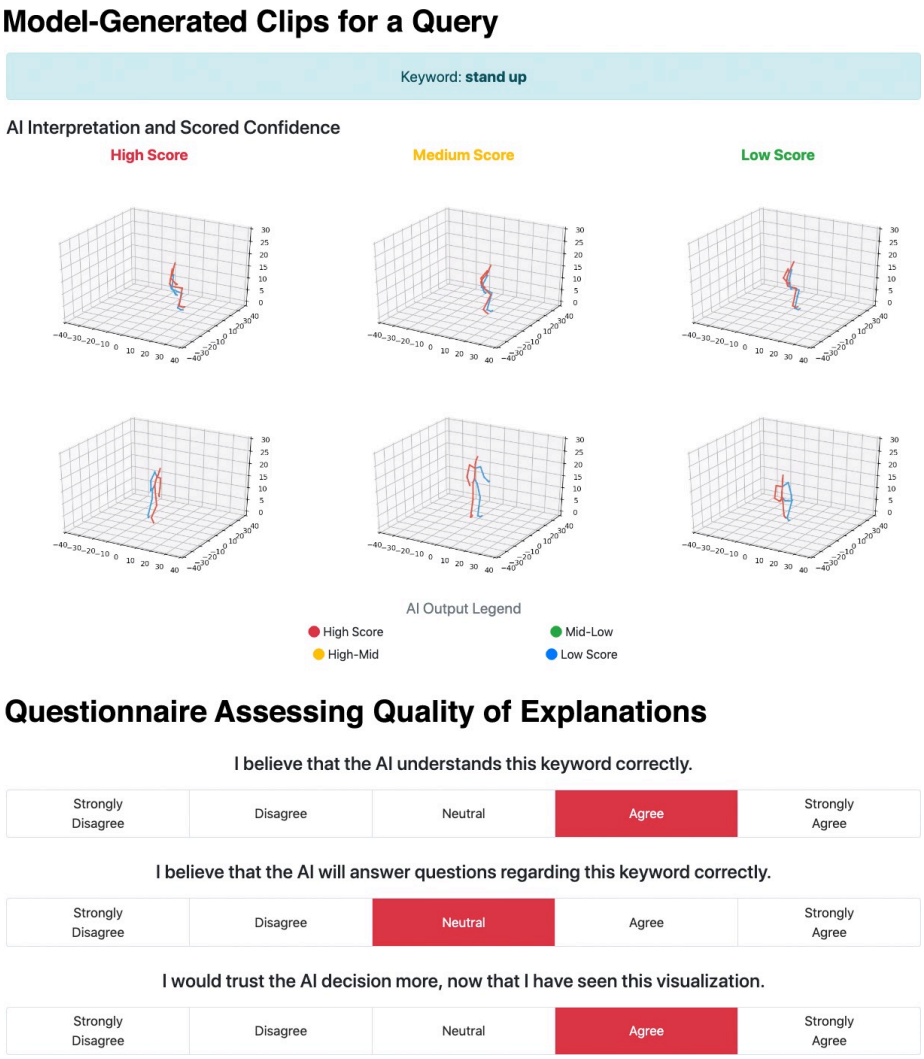

Figure 7.3: The XAI evidence screen. This is the first step towards assessing the mental model of the user. Before each task, the participant is presented with a query and a sorted list of generated clips depicting what the XAI thinks the query visually looks like. The participant is then prompted to answer a brief survey about their expectations of the XAI's ability to answer correctly and its understanding of the query.

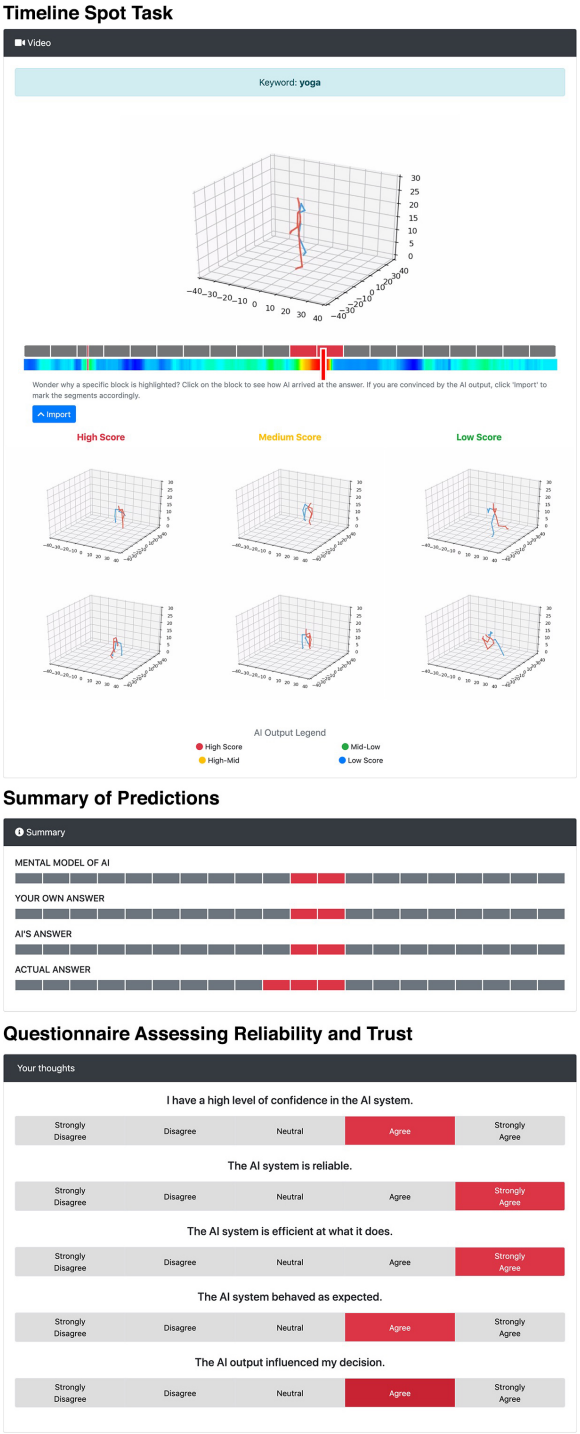

Figure 7.4: (top) The *Timeline Spot* task with the XAI system. Given a long video (with time indicator), the user's task is to manually highlight segments where the activity described by the given keyword query exists, or "import" AI assistance. The XAI system, unlike the AI system, provides assistance with explanations using model-generated clips. (middle) A summary of timelines with the location of the correct answer, the user's answer, the user's prediction of the system's answer, and the system's answer. (bottom) A questionnaire per trial assessing the performance of the system.

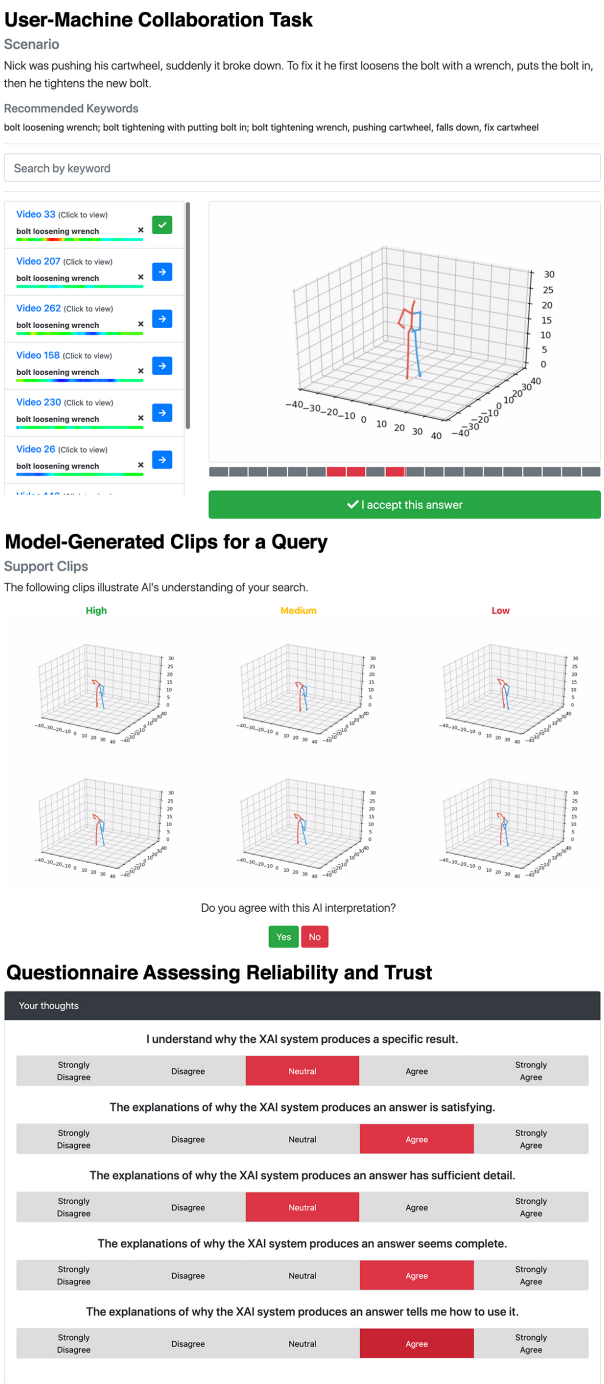

Figure 7.5: (top) The *User-Machine Collaboration* task with the XAI system. Given a series of long video clips and a more complex scenario, the user's task is to select a video that best represents the text description and manually curated hint keywords. The interface features a search box that allows the user to consult the (X)AI system to facilitate the investigation. (middle) The XAI system, unlike the AI system, provides assistance with explanations using model-generated support clips. (bottom) A questionnaire per trial assessing the performance of the system.

Part 2: Timeline Spot. The second part of the study provides a single but lengthier video clip, consisting of multiple individual clips as shown in Part 1, to localize a specific activity. Presenting a single keyword in the same fashion as the first part, the study prompts the participant to search for different parts of the video that best illustrate the keyword. The user can play or scrub the video to locate the parts that match the keyword, and mark them using the timeline interface as shown in Figure 7.4 (top). AI assistance is once again available for the user to consult, complete with AI-generated clips exclusive to the XAI system. Once the user clicks on the confidence bar below the clip, the supporting evidence consisting of generated clips is sorted automatically to show the most contributing evidence to a specific time segment. The trial structure mirrors the Clip Identify task, with a summary of results as shown in Figure 7.4 (middle), and a questionnaire as shown in Figure 7.4 (bottom).

Part 3: User-Machine Collaboration Task. Combining interface elements and challenges of parts 1 and 2, the third and final part of the study provides a large set of lengthier video clips and prompts a randomly selected scenario. Each of the seven scenarios was manually constructed by concatenating previously available motion-capture clips, and the list of distractor clips was compiled by random retrieval from the dataset. The user is encouraged to deconstruct the provided description and search for the video clip that best illustrates the scenario, but we suspect the task will be overwhelming enough for the user to request AI assistance as required. Illustrated in Figure 7.5, the interface provides a total set of three main elements: the search box, the clip list, and the video player. The user can independently browse and investigate the individual video clips to complete the task, but is encouraged to use the AI system to facilitate the investigation. Upon submitting one or more text queries, the AI system will highlight the clips that are most likely to illustrate the user's query. The XAI system, in alignment with its behavior in parts 1 and 2, presents its interpretation of the query through generated supporting evidence before the user accepts AI assistance in sorting the video clips. Finally, the user must continue the investigation until the correct video clip is selected, and then is

presented a summary of results and a questionnaire.

7.3.3 External Data Annotation

In addition to collecting participant reactions to model-generated video clips as part of the study, we recruited five external data annotators and launched a posthoc analysis of AI explanation quality. Each annotator was presented a series of keyword queries and corresponding AI explanations, and asked to rate how well the model-generated video clips represent each query, on a 5 point Likert scale. These ratings, collected from annotators with no prior experience with the study, were used as a proxy for the quality of AI explanations as well as an indicator of participant attentiveness throughout the study.

7.4 Outcomes

Participant activities recorded during each session have been collated and thoroughly analyzed to test the original hypothesis that XAI support will facilitate the user's understanding of the AI system and in turn improve the user's task performance. Any other notable insights that arose during this process have been also been collected for discussion below.

Measures. With a total of 44 participants engaging in more than 350 distinct AI-assisted tasks, the collected data features completion time, user and AI accuracy, and user reaction to system-provided explanations as applicable per task. In the following discussion, we note statistical tests with * at $p < .05$ and ** at $p < .005$.

Task Clusters. Upon observing divergence in participant performance and reaction between those who explicitly stated low levels of trust in AI explanation and those who did not, the study results were further segmented into three separate groups: AI tasks (51.7%), XAI tasks completed by users with low levels of trust (XAI LOW, 16.5%), and finally, the remainder of XAI tasks where the user did not express explicit distrust or instead expressed trust (XAI

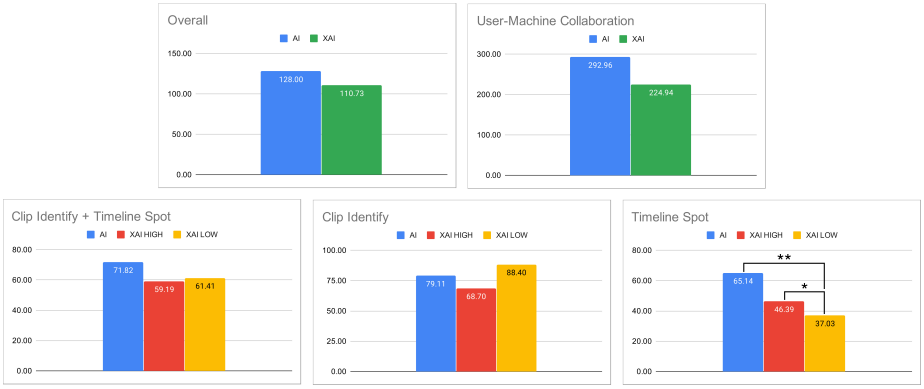

Figure 7.6: Task completion times (in seconds) between the two AI systems, and across the three task clusters. Segmentation between XAI HIGH and XAI LOW was determined by the user's express level of trust in AI explanation. UMC completion times using the XAI system remain unsegmented, as users did not provide express level of trust for corresponding AI explanation.

HIGH, 31.8%). Segmentation between XAI LOW and XAI HIGH clusters was determined by user response to the question "I would trust the AI decision more, now that I have seen this visualization," where the "Disagree" or "Strongly Disagree" response serving as a qualifier for XAI LOW. Some task trials were discarded due to user or system error, resulting in a slight imbalance between AI and XAI task numbers. A number of participants differed in each condition, though the results were segmented by task, not participants.

7.4.1 Speed

Overview. Speed is defined as the elapsed time in completing a single task trial, illustrated in Figure 7.6. Speed determines the efficiency advantage of using the AI or XAI system. Adjusting for variance in internal loading time for both AI and XAI systems, a typical task was completed on average in 119 seconds, although it is important to note that User-Machine Collaboration (UMC) tasks are more complex and hence more time-consuming for users. Excluding these collaborative tasks that require about 257 seconds to complete on average, the average completion time hovered around 66 seconds. Computation time was offset to allow for direct comparison between AI and XAI systems after the

study, although we recognize that the users may have deemed computation time excessive and influential to user satisfaction with the system.

Results Without task segmentation, the XAI system (111 seconds) presented negligible advantage over the AI counterpart (128 seconds), but more significant divergence emerged upon segmenting the XAI results by trust and task types. An ANOVA revealed no significant effect of task cluster on speed for Clip Identify (AI: M=79s, SD=78s, XAI HIGH: M=69s, SD=58s, XAI LOW: M=88s, SD=70s). Similarly, no significant effect of condition was found for the UMC tasks (AI: M=293s, SD=206s, XAI: M=225s, SD=164s). However, the Timeline Spot tasks varied significantly (**, $F(3,126)=5.21$, $p = .007$) with XAI LOW tasks being completed most quickly (M=37s, SD=39s), followed by XAI HIGH Tasks (M=46s, SD=36s) and AI tasks (M=65s, SD=46s). Post-hoc pairwise t-tests with Bonferroni correction for repeated measures revealed significant differences in completion time between AI and XAI HIGH (*, $p = .03$) and between AI and XAI LOW (**, $p = .002$). There was no significant difference between XAI HIGH and XAI LOW.

Discussion The provision of XAI support did not aid in the speed of task completion for the Identify task, as participants generally viewed multiple clips in detail, irrespective of XAI support. In the Timeline Spot task, overall completion times were generally shorter than either UMC or Identify counterparts, indicating a simpler task overall: in case of the XAI system, participants could use the AI explanation to know quickly whether to accept the AI answer or at least seek the playback to the highest rated positions to check them. The UMC task was designed as complex challenge which would maximize the support provided to participants, the results were not significantly different between the two systems, likely due to the very high variance between participants on the time to complete this task. This points to the individualized nature of the provision of evidence, and that it may be important to provide support on demand, while putting potentially distracting explanations out of the way when they are not requested or required.

7.4.2 Accuracy

Overview Accuracy, depicted in Figure 7.7, is the portion of instances where the user, with the help of the AI system, was able to identify the correct answer in a single task trial. Accuracy determines whether the system is able to produce more correct answers than others, resulting in a less error-prone experience.

Results The accuracy was highest for the XAI HIGH cluster (74.0%), followed by AI (68.2%) and XAI LOW (44.4%). Pairwise chi-square tests with Bonferroni correction revealed significant differences between AI and XAI LOW (**, $\chi^2(1,189) = 9.14$, $p = .002$) and between XAI HIGH and XAI LOW (**, $\chi^2(1,158) = 13.50$, $p = .0002$). The difference between AI and XAI HIGH was not significant.

Discussion The accuracy results were lowest when users indicated low trust for AI explanations. This may indicate that when trust is low, the user may assume the generated evidence is unreliable, and proceeded to submit their own (often incorrect) answer. This conjecture is reinforced by the fact that when trust is low, affecting accuracy, synchronization is usually also low. These results, for clarity, entirely depend on each users interaction with the system, independent of the underlying algorithm: one can choose to accept or ignore AI assistance, regardless of the system type in use.

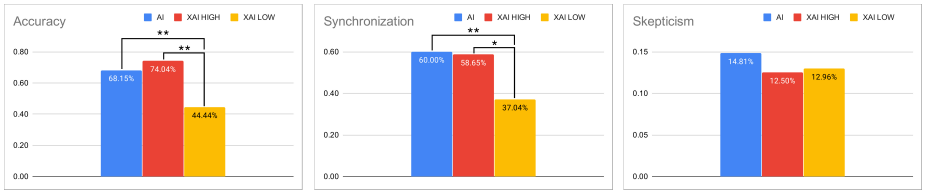

Figure 7.7: Results of accuracy (left), user-machine synchronization (middle), and user skepticism (right) across the three task clusters. Segmentation between XAI HIGH and XAI LOW was determined by the users' expressed level of trust in each AI explanation in "Clip Identify" and "Timeline Spot."

Figure 7.8: Truncated summary of participant "journeys" through the study across different experimental conditions. Each section displays two of the most successful journeys (where both the user and the AI system were able to identify the correct answer), as well as two of the least successful per corresponding condition. Top results indicate that both the user and the AI system were able to get correct answers throughout the journey, achieving a high level of synchronization.

7.4.3 User-Machine Synchronization

Overview Determined as the instance where both the user and the AI systems select the same answer regardless of its accuracy, this measure represented in Figure 7.7 defines the level of synchronization between the user and the AI system. As a whole, about 55% of all user and AI answers were synchronized. A sample of these journeys is illustrated in Figure 7.8. The icons indicate the user's accuracy and synchronization with the AI system's interpretation, and the numbers below indicate the user's understanding, confidence, and trust pertaining to the XAI system on a 1-5 scale. Queries used in each trial are also displayed, with colors indicating the ratings by external data annotators (green-yellow-red, in order by external rating). The full set is available in Chapter 9.

Results In strong alignment with the previous accuracy results, XAI LOW tasks resulted in a significantly lower synchronization rate of 37.40% in comparison to AI (60.0%) and XAI HIGH (58.65%) tasks. Post-hoc chi-square tests with Bonferroni correction revealed significant differences between AI and XAI LOW (**, $\chi^2(1,189) = 8.17$, $p = .004$) and between XAI HIGH and XAI LOW (*, $\chi^2(1,158) = 6.65$, $p = .0099$). The difference between AI and XAI HIGH was not significant.

Discussion AI and XAI HIGH results indicate higher user-machine synchronization than XAI LOW. This may indicate that the provision of trustworthy evidence (XAI HIGH) does not help any more than no evidence (AI), but the provision of untrustworthy evidence, such as poorly generated clips (XAI LOW) can actually drive participants away from AI suggestions. This is in fact the desired result, as we hope that users will appropriately choose to find their own answers when they do not trust the AI system to do the job.

7.4.4 User Skepticism

Overview When the user decides that the AI system's assistance is unhelpful and even incorrect, the user may explicitly exhibit a level of skepticism, illustrated in Figure 7.7, by choosing a correct answer despite the AI system's invalid

suggestion. About 14% of all tasks reflected this rare but consistent behaviour.

Results There was no significant deviation to trend across the three clusters, with AI, XAI HIGH, and XAI LOW tasks exhibiting evidence of skepticism 14.8%, 12.5%, and 13.0% at a time respectively.

Discussion User skepticism serves as a proxy measure of user attention to task, indicating that the participants sometimes went against AI suggestions and did not blindly accept them. This phenomenon was consistent across all task clusters, and there was no correlation between this behaviour and expert ratings per clip.

Question	AI System	XAI System High trust	XAI System Low trust
MAIN_Q1 I have a high level of confidence in the AI system.	3.37	3.95	3.02
MAIN_Q2 The AI system is reliable.	3.42	4.23	3.00
MAIN_Q3 The AI system is efficient at what it does	3.46	4.25	3.14
MAIN_Q4 The AI system behaved as expected.	3.51	4.27	3.23
MAIN_Q5 The AI output influenced my decision.	2.93	3.94	2.69
MENTAL_Q1 I have a clear understanding of how AI would answer this question.	3.71	4.09	3.09
MENTAL_Q2 I am confident that my answer would match the AI interpretation.	3.71	4.28	3.03
END_Q1 I understand why the AI system produces a specific result.	3.86	4.10	3.26
END_Q2 The explanations of why the AI system produces an answer is satisfying.	3.77	4.40	3.15
END_Q3 The explanations of why the AI system produces an answer has sufficient detail.	3.59	4.26	2.91
END_Q4 The explanations of why the AI system produces an answer seems complete.	3.63	4.15	3.06
END_Q5 The explanations of why the AI system produces an answer tells me how to use it.	3.80	4.15	3.03

Figure 7.9: Summary of questionnaire responses about the (X)AI system, listing individual questions featured in the original study session. Responses pertaining to the XAI system are split into two parts, based on the overall level of trust and reliance indicated by each study participant, represented by individual responses to questions pertaining to model-generated clips. If the user expressed general lack of confidence in AI explanation, the user was classified as "low trust." The XAI system performed better than the AI system without explanation, if the user had a higher level of trust in the system.

7.4.5 Questionnaire Responses

Observation While the AI-only system originally seemed to yield higher overall satisfaction amongst the participants, there was a sharp divide in satisfaction between the participants with a high level of trust and reliance for the XAI system compared to those without. Upon segmenting the responses from the XAI system as illustrated in Figure 7.9, it was evident that the XAI system resulted in a more positive experience overall compared to the AI system, should the users have a high level of trust and reliance for the system.

7.4.6 Additional Findings

Overview Below are some of the secondary findings that do not directly correspond our hypothesis, but are notable and warrant further investigation in future work.

Distribution of User Reactions to AI Explanation The user study collected using the three distinct questions, to individual AI explanations: "I believe that the AI understands this keyword correctly" (UNDERSTAND), "I have a high level of confidence in the AI system" (CONFIDENCE), and "I would trust the AI decision more, now that I have seen this visualization" (TRUST). Upon visualizing these reactions, there was apparent bimodal behaviour, as illustrated in Figure 7.10, across all three categories, indicating that users often exhibit less ambiguous reactions to presented AI explanations. We recognize that these reactions are biased to each participants subjective experience.

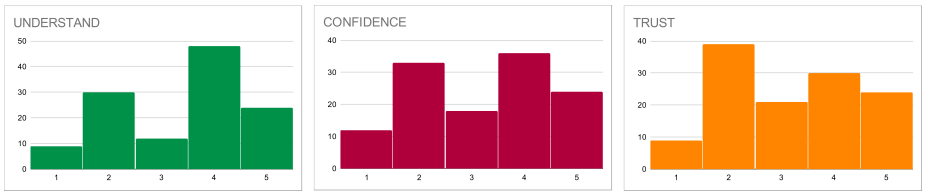

Figure 7.10: Histograms illustrating user reactions to AI explanation, represented by responses to three distinct questions about AI assistance. The pattern shows a bimodal distribution showing that participants formed clear opinions for most task trials, especially on the understand and confidence questions.

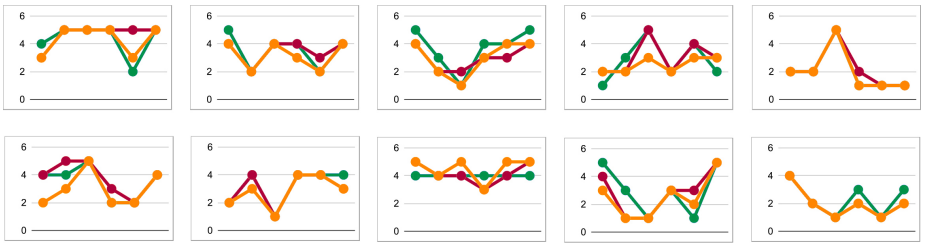

Figure 7.11: Small multiples illustrating example reactions to AI explanations when using the XAI system, represented by three distinct questions: UNDERSTAND (green), CONFIDENCE (maroon), and TRUST (orange). Each plot represents a single participant across the tasks completed in the XAI experimental condition.

Correlation Between User Reactions to AI Explanation Beyond the anecdotal tendency where individual users who exhibit trust in the AI system may also indicate confidence in the AI system, as illustrated in Figure 7.11, there was significant correlation between the user's three responses to a specific AI explanation. Post-hoc multiple correlation tests revealed significant positive correlation across the board: UNDERSTAND and CONFIDENCE (**, $r(121) = 0.7979$, $p < .00001$), UNDERSTAND and TRUST (**, $r(121) = 0.7777$, p

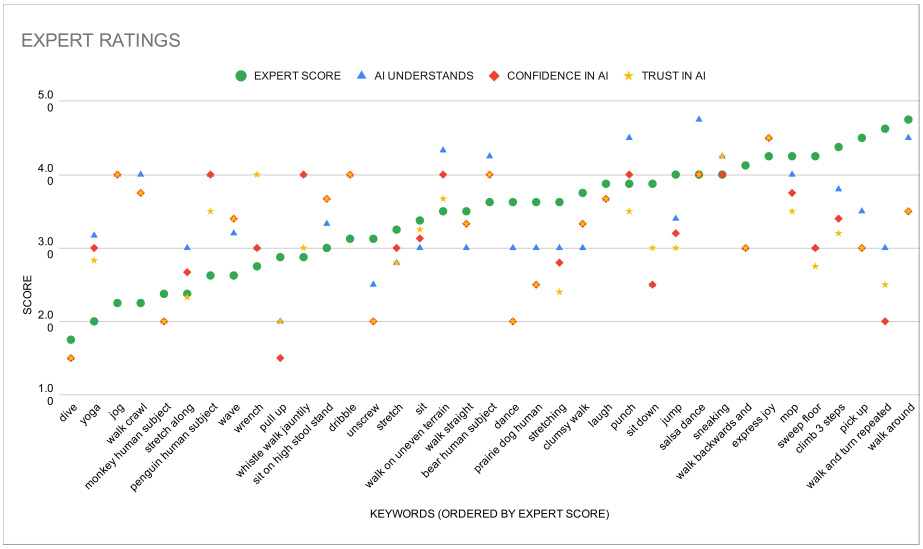

Figure 7.12: Alignment between external ratings and individual user reactions to AI explanations. X axis is anchored by keywords ordered by external rating scores.

<.00001), and CONFIDENCE and TRUST (**, $r(121) = 0.7777$, $p <.00001$). This, along with the shifting user reactions in Figure 7.11, suggest that users do actively respond to presented AI explanation and change their opinions accordingly, and that not all three questions may be necessary in future studies to measure the level of trust and reliance on the system.

Alignment with External Ratings There was no apparent correlation between user reactions to individual clips and externally annotated ratings, as indicated in Figure 7.12. This may indicate that clip quality assessment criteria differed between our experts and participants, or that overall clip quality did not strongly influence the user's trust or confidence in the AI agent. It was notable, however, that positive user reactions were clustered around clips that feature exaggerated motions and cartoon-like premises, such as "bear (human subject)," "salsa dance," and "express joy," while more generic and muted clips such as "pull up" and "walk and turn repeated" received negative reactions.

Participant Comments There was divergence between clusters of participants who found the AI system to be reliable and influential to their decision-making processes, and those who deemed the system to be counter-intuitive and underwhelming. One participant wrote "(AI explanation) is a good basis of determining the reliability of AI in terms of (whether) the AI is able to detect the proper animations," and another expressed satisfaction, stating "(I am) impressed of what the AI system outputs." On the other hand, some expressed caution and distrust, with one writing "the AI system often interpreted small portions of movements as if they met the definition of the keyword although it was a mere segment of the movement," and another writing "I didn't trust it completely as it directed similar movements and categorized it as the real one." Two participants plainly wrote "I did not see the explanation," alluding to the possibility that the definition or qualifications of what constitutes an AI *explanation* may vary between individuals or may require additional training or clearer messaging to help people interpret generated clips as explanations.

7.5 Discussion

Outcomes The user study outcomes present significant evidence that the XAI system and its generative examples can facilitate task performance consistent with the AI system, offer improved performance in select task types, and provide a more satisfying overall user experience. However, this is only applicable if the users decide to trust the provided AI explanations. To clarify, this study is not designed to influence the user's reaction to a certain system, but to provide additional tools for further clarity and transparency and observe how one perceives AI-provided explanation.

We claim that the presence of AI explanations, characterized by exemplar clips and the corresponding interactive visualization, does not improve the users performance in search tasks, but helps one to know when to trust or reject AI assistance, thus indirectly influencing performance. Additionally, the presence of such visualization helps to identify the user as belonging in one of the two groups: those who exhibit a higher level of trust and satisfaction for the system, and those exhibit skepticism and yield a lower level of efficiency and accuracy.

We observed a significant divide in behavior and performance between users who chose to trust the AI explanations and those who did not, and this divide impacted all performance-related measures including speed, accuracy, and user-machine synchronization. While there was no significant indication that the user was able to correctly accept or reject the XAI system's assistance, the results were largely comparable with the AI counterpart. These results suggest that users form trust and affinity for the XAI or AI system more or less based on instinct, and the system may produce video clips that ultimately result in correct answers, but not necessarily seem logical or comprehensible to human users. This disparity contributes to lack of perceived performance improvement.

These findings are consistent with other study results that examine participant interaction with interpretable AI models: trust levels are found to be inferred by subjective perceptions of the machine performance [166]; model explanations are found to offer no significant improvement to user performance [124]; finally, one's personal characteristics significantly influence perception of AI systems, regardless of their objective performance [108]. Such alignment

also presents an opportunity to study how high-stake, ambiguous situations may alter human-AI dynamics where experts may strongly disagree on existing assessments and need to rely on a tie-breaking nudge by an AI system with a level of explainability [141].

Future Improvement. The XAI system could be further improved in numerous areas to gain a more significant advantage over the AI-only counterpart. The generative model could be improved to produce more exemplars that achieve the same or higher level of accuracy as the AI-only system. Also, the XAI system could produce more high-quality model-generated clips that best represent individual queries and result in higher user satisfaction and user-machine synchronization. Explanations must be short and require little effort to interpret, or the advantage they offer will be outweighed by the required time and effort.

We also hypothesize that the evaluation dataset may contain human error and biases that may contribute to the system behaving in a way unexpected and even jarring to the users. For users to make more accurate, informed decisions, the system will need to transparently communicate what the potential biases are, and why its decision, although less intuitive, can result in the correct answer.

There are other implications pertaining to the experiment design as well. Users may exhibit a higher level of trust and reliance for the XAI system, should the individual tasks present a higher stake and a more captivating incentive. Task formulation is an important consideration as well: all the presented tasks in the study are simple permutations of the same dataset and the interface components, yet user performance and satisfaction noticeably differ across the tasks. Experimenting with different configurations may be useful in identifying user biases and designing tasks with more balanced challenges.

7.6 Summary

Built on a novel explainable approach for searching and ranking videos using textual queries, this chapter sets out to evaluate the XAI system's benefits over the black box counterpart in realms of overall explainability, trustworthiness, and accuracy. The following findings emerged from the featured user study:

- The XAI system yielded a comparable level of efficiency, accuracy, and user-machine synchronization as the AI system, but only if the user exhibited a high level of trust.
- The XAI system yielded a significantly lower level of efficiency, accuracy, and synchronization if the user exhibited a lower level of trust.
- The XAI system yielded higher overall user satisfaction, but only if the user exhibited a higher level of trust.

Chapter 8

Conclusion

We can't say who has come, perhaps we will never know, but many signs indicate that the future enters us in this way in order to be transformed in us, long before it happens ... just as people for a long time had a wrong idea about the sun's motion, they are even now wrong about the motion of what is to come. The future stands still, dear Mr. Kappus, but we move in infinite space.

Rainer Maria Rilke, *Letters to a Young Poet* [135]

8.1 Summary and Future Opportunities

Inspired by the increasing need for interface tools to work with complex language and image data together in a scalable fashion, Modular is a highly customizable, extensible software platform for visualizing available model outputs, building and annotating new datasets, and setting up user studies. This toolbox brings together the usually disparate actions of data annotation and curation, as well as machine learning visualization and testing. This toolbox enables vision and language researchers to seamlessly conduct their work without complex configuration of a web-interface, and furthermore invites other developers to build new modules per project needs. Other subsequent projects that rely on Modular further underscore its capabilities as a flexible and extensible foundation for human-in-the-loop AI research work.

8.1.1 Annotation for Commonsense Grounding

Armed with the goal of establishing a generalizable knowledge base for future AI work, Aesop presents a strong case in augmenting previous datasets with external detector outputs and user-provided annotations with a novel end-to-end grounding procedure. By facilitating rich media assets and deploying a flexible knowledge graph conducive to human augmentation and contextual expansion, Aesop demonstrates the benefits of encoding complex scene information to inform future creative projects or fuel complex AI models. In addition, Aesop further underscores the value of user-oriented annotation experience as the domain expert, initially unfamiliar with annotation tasks, was able to ground the previously disconnected datasets and annotate the entire film with minimal technical support. The domain expert's annotation process accelerated over the subsequent scenes as the annotator became more familiar with the workflow.

The knowledge graph powering Aesop has the potential to include representations of other feature-length films to form a larger and richer body of film knowledge. While the annotation work presented in the previous chapter is limited to a single feature film whose initial datasets and detector outputs are available for grounding, this project can be greatly expanded upon to invite more human annotators absorb future datasets. In addition to accumulating a complex network of domain-specific insights ranging from narrative patterns to cinematography techniques, Aesop will be able to establish a much richer digitized environment and complex knowledge/belief model that bear resemblance to their source films, many of which are in turn based on the real world.

8.1.2 Human-AI Collaboration in Content Creation Tasks

The Modular framework was adapted to facilitate human-AI collaboration in the context of content creation via two distinct use cases. Displaying its ability to flexibly customize its design and functionality per project-specific needs, the framework demonstrates a range of human-AI collaborative tasks such as 3D animation generation and music composition with AI assistance. Using Aesop's conversational AI mode and its knowledge graph accumulated through the commonsense grounding process, one can produce new and original film scenes using

natural language parsing, human gesture integration, and chat interface. With MUSICA's web-based interface, musicians can seamlessly compose and improvise music with their AI partners by combining domain-specific representation framework, natural language processing, and real-time music generation.

Beyond the presented use cases, the platform can be flexibly applied to other projects that feature collaborative tasks between human users and AI models, especially should convenience in deployment and user access be important to the end product: Modular's cloud-hosted, standard-compliant interface is advantageous in rapid deployment and ensuring that users can access the web application without installing native software. In addition to closely-coordinated research collaborations, Modular can benefit from additional software engineering work to serve the general public as a self-serve platform: this idea has been already explored with the previously presented layout generator, and will allow all interested users to conveniently build and deploy an interface without development assistance.

8.1.3 Approach to Improved AI Explanation

With transparency and interpretability emerging as increasingly important considerations in user-facing AI solutions, the generative XAI system presented in the previous chapter sets out to provide AI assistance that can explain its decision-making process without compromising system performance. Conventional AI models are often considered "black-box" for the lack of clarity in their inner workings, and the visualizations that accompany them often are focused on illustrating learned representations and influential input features, instead of helping non-experts to understand the system. Inspired by the nature of GAN models where the generator component dynamically creates visual assets to use as queries for the discriminator component, the generative XAI system uses the aforementioned video queries as partial explanations of the AI model.

The concept of using the products of adversarial generation as explanation elements can benefit from further exploration in different media types and application areas. Beyond skeletal 3D animations presented in the previous chapter, other AI systems may experiment with presenting various interim iterations of

text, image, or audio generation to the users in a manner conducive to user interpretation. In addition to improving explainability of the AI system, these assets may be directly useful in the user's workflow. For instance, the user may be able to identify alternate queries upon observing the explanations and gaining a better understanding of the AI system in one's search task. On the other hand, the user may be able to derive additional sources of inspiration from these by-products of GAN models when working on a creative project.

8.1.4 Trust and Reliance in Human-AI Collaboration

Built on the previously presented approach to AI explanation via visual queries, the XAI system was compared against the black box counterpart to evaluate its benefits using different criteria and its ability to establish the user's trust and reliance on the system. As the study results suggest, trust was not a mere product of the XAI system's overall performance, but a factor in the successful cooperation between the human user and the system: the XAI system resulted in a comparable level of efficiency, accuracy, and user-machine synchronization as the AI system, but only if the user exhibited a high level of trust, and the same system yielded higher overall user satisfaction as an added effect as well.

As discussed in the previous chapter, the XAI system can benefit from additional evaluation with improved visual queries and system performance to conclude whether the XAI system can gain a significant advantage over the AI counterpart. Instead of simply presenting a generic set of keywords and tasks, the experiment could benefit from additional context — including more immersive objectives and captivating incentives — to observe how the user's trust and reliance on the system may change according to the nature of the interaction.

8.1.5 Future Opportunities

Additional Modules. With emerging Web standards and innovative peripheral devices, such as consumer VR headsets, entering consumer markets, Modular continues to receive additional project-specific modules that take advantage of such technologies. For example, a graph module, which can alternatively used to indicate bird's eye view of multiple objects placed on a two-dimensional

plane, can be augmented with a new 3D space module that interprets and visualizes the individual objects in a three-dimensional space: coupled with head-mounted display devices such as Oculus Rift, the result is an immersive, first-person experience achieved without the use of dedicated software and a lengthy development cycle that may follow. In addition, the rising demand for real-time communication capabilities can be readily satisfied with text chat and "co-presence" modules that can turn an otherwise solitary annotation or visualization experience into a collaborative, team-based one. Finally, anecdotal evidence derived from experience with Modular reaffirms that utilizing modules, whose core features can be reused and extended, offer opportunities for productivity gains as supported by significant evidence in the software industry [110]. Equipped with a suite of core modules that perform basic yet specific functions, Modular allows its developers to identify the modules that can be reused and bridge feature gaps with further customization or extension. This approach, albeit anecdotally, resulted in an accelerated development and deployment rate as researchers became more familiar with Modular's original offerings and started to accumulate project-specific modules, one project at a time.

Crowdsourcing Platform. Having identified a market gap between survey-focused online services (e.g., SurveyMonkey) and research-oriented crowdsourcing platforms (e.g., Amazon Mechanical Turk), Modular may serve as an attractive alternative between these product categories for those who wish to conduct more complex crowdsourced studies without the steep learning curve involved in developing custom software from scratch. In addition to essential features found in typical survey tools, Modular can promote more complex annotation and visualization tasks with the layout generator and relevant modules — inspiring more users to construct and participate in crowdsourced research projects.

Collaboration in Creative Tasks. With Modular successfully deployed in research projects that feature scene annotation and collaborative music making, there remain future opportunities for exploring human-AI collaboration in other creative endeavours. As creative processes increasingly make their way onto the Web, partly due to shifting user demands and ubiquity of web technologies, Modular can potentially serve as a mechanism for researchers to easily perform

ethnographic research in different creative disciplines and study implicit user response to AI assistance without lab-based arrangements.

8.2 Concluding Remarks

Modular, true to its name, serves as a useful framework for taking a module-driven approach to satisfying a variety of project needs, and it does so by demonstrating its capabilities as an annotation interface, a collaborative workbench for human-AI tasks, and a rapidly deployed user study application. Modular is not simply limited to such use cases, though: as one can construct new and innovative interfaces with its layout generator and user-defined modules, taking cues from conventional website builders to aid in improving the workflow of other research effort.

Though seemingly as old as time, humans and machines continue to maintain a close yet tense relationship as AI quickly ramps up its capabilities with machine learning, computer vision, and natural language processing. AI and its creators face increased ethical scrutiny surrounding personal information and data privacy; human users continue to offer their insights and information in exchange for convenience, all while comparing AI to a mystical black box. There remains an opportunity to build a world where humans and machines collaborate beyond simply coexisting, and Modular stands a promising platform for facilitating future XAI research, with adaptability and customizability as its key strengths.

Bibliography

[1] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[2] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh. Text2Action: Generative Adversarial Synthesis from Language to Action. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5, 2018.

[3] J. Allen and G. Ferguson. Human-machine collaborative planning. In *Third International NASA Workshop on Planning and Scheduling for Space*, pages 27–29, 2002.

[4] J. F. Allen, M. Swift, and W. de Beaumont. Deep Semantic Analysis of Text. In *Conference on Semantics in Text Processing*, Step 08. Association for Computational Linguistics, 2008.

[5] F. S. an Dmitry Kalenichenko and J. Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. *Computing Research Repository*, abs/1503.03832, 2015.

[6] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[7] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, M. Eskevich, R. Ordelman, G. J. F. Jones, and B. Huet. TRECVID 2017: Evaluating Ad-hoc and Instance

Video Search, Events Detection, Video Captioning and Hyperlinking. In *TRECVID*. NIST, 2017.

[8] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, volume 33, pages 2429–2437, 2019.

[9] E. Barsoum, J. Kender, and Z. Liu. HP-GAN: Probabilistic 3D human motion prediction via GAN. *arXiv*, abs/1711.09561, 2017.

[10] R. Beale and C. Creed. Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies*, 67(9):755–776, 2009.

[11] C. Beattie, J. Z. Leibo, D. Teplyashin, T. Ward, M. Wainwright, H. Küttler, A. Lefrancq, S. Green, V. Valdés, A. Sadik, J. Schrittwieser, K. Anderson, S. York, M. Cant, A. Cain, A. Bolton, S. Gaffney, H. King, D. Hassabis, S. Legg, and S. Petersen. DeepMind Lab. *Computing Research Repository*, abs/1612.03801, 2016.

[12] A. L. Beberg, D. L. Ensign, G. Jayachandran, S. Khaliq, and V. S. Pande. FoldingHome: Lessons from eight years of volunteer distributed computing. In *IEEE International Symposium on Parallel & Distributed Processing*, pages 1–8. IEEE, 2009.

[13] J. M. Bradshaw, A. Acquisti, J. Allen, M. R. Breedy, L. Bunch, N. Chambers, P. Feltovich, L. Galescu, M. A. Goodrich, R. Jeffers, et al. Teamwork-centered autonomy for extended human-agent interaction in space applications. In *Association for the Advancement of Artificial Intelligence*, pages 22–24, 2004.

[14] R. Brath and E. Banissi. Using text in visualizations for micro/macro readings. In *IUI Workshop on Visual Text Analytics*. ACM, 2015.

[15] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 708–713. IEEE, 2005.

[16] R. O. Briggs, G.-J. De Vreede, and J. F. Nunamaker Jr. Collaboration engineering with ThinkLets to pursue sustained success with group support systems. *Journal of Management Information Systems*, 19(4):31–64, 2003.

[17] M. Butcher. *Building Websites with OpenCms*. Packt Publishing Ltd, 2004.

[18] C. J. Cai, J. Jongejan, and J. Holbrook. The effects of example-based explanations in a machine learning interface. In *International Conference on Intelligent User Interfaces*, pages 258–262. ACM, 2019.

[19] M. Callier and H. Callier. Blame It on the Machine: A Socio-Legal Analysis of Liability in an AI World. *Washington Journal of Law, Technology & Arts*, 14:49, 2018.

[20] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *Computing Research Repository*, abs/1611.08050, 2016.

[21] J. C. Chang, S. Amershi, and E. Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 2334–2346, 2017.

[22] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han. Show, Observe and Tell: Attribute-driven Attention Model for Image Captioning. In *International Joint Conference on AI*, 2018.

[23] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *Computing Research Repository*, abs/1512.01274, 2015.

[24] J. Choo and S. Liu. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications*, 38(4):84–92, 2018.

[25] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and Trust: Designing Model-driven Visualizations for Text Analysis. In *SIGCHI Conference on Human Factors in Computing Systems*, 2012.

[26] D. Coutu and M. Beschloss. Why teams don't work. *Harvard Business Review*, 87(5):98–105, 2009.

[27] P. C. F. da Costa. Conversing with personal digital assistants: on gender and artificial intelligence. *Journal of Science and Technology of the Arts*, 10(3):2–59, 2018.

[28] T. H. Davenport and R. Ronanki. Artificial intelligence for the real world. *Harvard Business Review*, 96(1):108–116, 2018.

[29] E. Davis and G. Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

[30] J. de Kleer and R. Reiter. Foundations for assumption-based truth maintenance systems: Preliminary report. In *Association for the Advancement of Artificial Intelligence*, pages 183–188, 1987.

[31] C. Decurtins, M. C. Norrie, and B. Signer. Digital annotation of printed documents. In *Conference on Information and Knowledge Management*, pages 552–555, 2003.

[32] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel. The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. In *Hawaii International Conference on System Sciences*, 2019.

[33] S. Deterding, J. Hook, R. Fiebrink, M. Gillies, J. Gow, M. Akten, G. Smith, A. Liapis, and K. Compton. Mixed-initiative creative interfaces. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 628–635, 2017.

[34] M. d'Inverno and J. McCormack. Heroic versus collaborative AI for the arts. *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2015.

[35] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan. Explaining models: An empirical study of how explanations impact fairness judgment. *arXiv preprint arXiv:1901.07694*, 2019.

[36] R. Eckart de Castilho, E. Mujdricza-Maydt, S. Yimam, S. Hartmann, I. Gurevych, A. Frank, and C. Biemann. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *LT4DH Workshop at COLING*, 2016.

[37] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. Riedl. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. *arXiv preprint arXiv:1901.03729*, 2019.

[38] M. Eiband, S. T. Völkel, D. Buschek, S. Cook, and H. Hussmann. When people and algorithms meet: user-reported problems in intelligent everyday applications. In *International Conference on Intelligent User Interfaces*, pages 96–106. ACM, 2019.

[39] H. Ekbia and B. Nardi. Heteromation and its (dis) contents: The invisible division of labor between humans and machines. *First Monday*, 2014.

[40] M. El-Assady, V. Gold, C. Acevedo, C. Collins, and D. Keim. ConToVi: Multi-Party Conversation Exploration using Topic-Space Views. In *Computer Graphics Forum*, volume 35, pages 431–440, 2016.

[41] A. C. Elkins, N. E. Dunbar, B. Adame, and J. F. Nunamaker. Are users threatened by credibility assessment systems? *Journal of Management Information Systems*, 29(4):249–262, 2013.

[42] M. Erdmann, A. Maedche, H.-P. Schnurr, and S. Staab. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. In *COLING Workshop on Semantic Annotation and Intelligent Content*, pages 79–85, 2000.

[43] D. G. et al. HARVEST: An Intelligent Visual Analytic Tool for the Masses. In *International Workshop on Intelligent Visual Interfaces for Text Analysis*, 2010.

[44] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. VSE++: improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017.

[45] J. Falk, S. Poulakos, M. Kapadia, and R. W. Sumner. PICA: Proactive Intelligent Conversational Agent for Interactive Narrative. *International Conference on Intelligent Virtual Agents*, 2018.

[46] F. Ferraro, N. Mostafazadeh, T.-H. Huang, L. Vanderwende, J. Devlin, M. Galley, and M. Mitchell. A Survey of Current Datasets for Vision and Language Research. In *Conference on Empirical Methods in Natural Language Processing*, pages 207–213, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[47] T. Fountaine, B. McCarthy, and T. Saleh. Building the AI-powered organization. *Harvard Business Review*, 97(4):62–73, 2019.

[48] K. Fragkiadaki, S. Levine, and J. Malik. Recurrent Network Models for Kinematic Tracking. In *Computer Vision and Pattern Recognition*, 2015.

[49] J. Gauthier. Conditional Generative Adversarial Nets for Face Generation. Technical report, Stanford, 2014. CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester.

[50] P. Ghosh, J. Song, E. Aksan, and O. Hilliges. Learning human motion models for long-term predictions. In *Computer Vision and Pattern Recognition*, 2017.

[51] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. *Computing Research Repository*, abs/1609.03677, 2016.

[52] J. Goetz, S. Kiesler, and A. Powers. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Workshop on Robot and Human Interactive Communication*, pages 55–60. IEEE, 2003.

[53] O. Gomez, S. Holter, J. Yuan, and E. Bertini. ViCE: visual counterfactual explanations for machine learning models. In *International Conference on Intelligent User Interfaces*, pages 531–535, 2020.

[54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[55] R. Guidotti, A. Monreale, and D. Pedreschi. The AI black box explanation problem. *ERCIM News*, 116:12–13, 2019.

[56] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANS. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[57] D. Gunning. Explainable Artificial Intelligence. Technical Report Darpa-baa-16-53, Darpa, 2016.

[58] B. Hahn. CMU Graphics Lab Motion Capture Database Motionbuilder-friendly BVH conversion. `https://sites.google.com/a/cgspeed.com/cgspeed/motion-capture/cmu-bvh-conversion`, Accessed 2018.

[59] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016.

[60] K. M. Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin, M. Wainwright, C. Apps, D. Hassabis, and P. Blunsom. Grounded Language Learning in a Simulated 3D World. *Computing Research Repository*, abs/1706.06551, 2017.

[61] A. Hertzmann. Can computers create art? In *Arts*, volume 7, page 18. Multidisciplinary Digital Publishing Institute, 2018.

[62] D. Hewlett, S. Hoversten, W. Kerr, P. R. Cohen, and Y.-H. Chang. Wubble World. In *Artificial Intelligence for Interactive Digital Entertainment*, pages 20–24, 2007.

[63] I. Higgins, A. P. Loic Matthey, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. Beta-VAE: Learning Basic Visual Concepts with a

Constrained Variational Framework. In *International Conference on Learning Representations*, 2017.

[64] E. A. Holm. In defense of the black box. *Science*, 364(6435):26–27, 2019.

[65] S. R. Hong, J. Hullman, and E. Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *ACM Conference on Computer Supported Cooperative Work (CSCW)*, 2020.

[66] J. Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6):1–4, 2006.

[67] P. Hudak, D. Quick, M. Santolucito, and D. Winograd-Cort. Real-time interactive music in haskell. In *ACM SIGPLAN International Workshop on Functional Art, Music, Modelling and Design*, pages 15–16, 2015.

[68] S. Ibarluzea. Annotorious - Image Annotation for the Web. https://universaldatatool.com/, 2020. [Online; accessed 25-March-2020].

[69] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *Computer Vision and Pattern Recognition*, 2016.

[70] M. H. Jarrahi. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586, 2018.

[71] M. Johnson, K. Hofmann, T. Hutton, and D. Bignell. The Malmo Platform for Artificial Intelligence Experimentation. In *International Joint Conference on Artificial Intelligence*, 2016.

[72] M. Kapadia, S. Frey, A. Shoulson, R. W. Sumner, and M. H. Gross. CANVAS: computer-assisted narrative animation synthesis. *Symposium on Computer Animation*, pages 199–209, 2016.

[73] A. Karpathy, J. Johnson, and L. Fei-Fei. Visualizing and understanding recurrent networks. 2015.

[74] C. Kennington, S. Kousidis, and D. Schlangen. Situated Incremental Natural Language Understanding using a Multimodal, Linguistically-driven Update Model. *International Conference on Computational Linguistics*, pages 1803–1812, 2014.

[75] W. Kerr, S. Hoversten, D. Hewlett, P. Cohen, and Y.-H. Chang. Learning in Wubble World. In *IEEE International Conference on Development and Learning*, pages 330–335. IEEE, 2007.

[76] S. Kim, D. Salter, L. DeLuccia, K. Son, M. R. Amer, and A. Tamrakar. SMILEE: Symmetric Multi-modal Interactions with Language-gesture Enabled (AI) Embodiment. In *North American Association for Computational Linguists*, 2018.

[77] P. Kingsbury and M. Palmer. From Treebank to Propbank. In *International Conference on Language Resources and Evaluation*, 2002.

[78] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302, 2015.

[79] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

[80] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. Personalized explanations for hybrid recommender systems. In *International Conference on Intelligent User Interfaces*, pages 379–390. ACM, 2019.

[81] J. Krause, A. Perer, and E. Bertini. Using visual analytics to interpret predictive machine learning models. *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2016.

[82] H. Lausberg and H. Sloetjes. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, 41(841):841–849, 2009.

[83] N. Lawson, K. Eustice, M. Perkowitz, and M. Yetisgen-Yildiz. Annotating large email datasets for named entity recognition with mechanical turk. In

NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 71–79, 2010.

[84] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[85] K.-F. Lee. A Blueprint for Coexistence with AI. *Wired Magazine*, Jul 2017.

[86] D. Li, P. P. Rau, and Y. Li. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, 2(2):175–186, 2010.

[87] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018.

[88] C. Liang, J. Proft, E. Andersen, and R. A. Knepper. Implicit communication of actionable information in human-AI teams. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[89] F. Liang, V. Das, N. Kostyuk, and M. M. Hussain. Constructing a data-driven society: China's social credit system as a state surveillance infrastructure. *Policy & Internet*, 10(4):415–453, 2018.

[90] A. Liapis, G. N. Yannakakis, C. Alexopoulos, and P. Lopes. Can computers foster human users' creativity? Theory and praxis of mixed-initiative co-creativity. *Digital Culture & Education*, 2016.

[91] O. S. C. V. Library. CVAT - Computer Vision Annotation Tool. https://github.com/opencv/cvat, 2018. [Online; accessed 25-June-2020].

[92] X. Lin and M. R. Amer. Human motion modeling using DVGANs. *arXiv preprint arXiv:1804.10652*, 2018.

[93] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards better analysis of deep convolutional neural networks. In *Transactions on Visualization and Computer Graphics*, volume 23, pages 91–100, 2016.

[94] L. Longdin. Liability for Defects in Bespoke Software: are Lawyers and Information Scientists Speaking the Same Language. *International Journal of Law and Information Technology*, 8:1, 2000.

[95] B. Lubars and C. Tan. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In *Advances in Neural Information Processing Systems*, pages 57–67, 2019.

[96] T. W. Malone. How human-computer 'Superminds' are redefining the future of work. *MIT Sloan Management Review*, 59(4):34–41, 2018.

[97] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep Extreme Cut: From Extreme Points to Object Segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[98] M. Marti, J. Vieli, W. Witoń, R. Sanghrajka, D. Inversini, D. Wotruba, I. Simo, S. Schriber, M. Kapadia, and M. Gross. CARDINAL: Computer assisted authoring of movie scripts. In *International Conference on Intelligent User Interfaces*, pages 509–519, 2018.

[99] J. Martinez, M. J. Black, and J. Romero. On human motion prediction using recurrent neural networks. In *Computer Vision and Pattern Recognition*, 2017.

[100] C. Matuszek. Grounded Language Learning: Where Robotics and NLP Meet. *International Joint Conference on Artificial Intelligence*, 2018.

[101] J. McCarthy et al. *Programs with Common Sense*. RLE and MIT computation center, 1960.

[102] J. McCormack, T. Gifford, P. Hutchings, M. T. Llano Rodriguez, M. Yee-King, and M. d'Inverno. In a silent way: Communication between AI and improvising musicians beyond sound. In *CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[103] M. McLuhan. *Understanding Media: The Extensions of Man*. 1964.

[104] T. Meo, A. Raghavan, D. A. Salter, A. Tozzo, A. Tamrakar, and M. R. Amer. Aesop: A Visual Storytelling Platform for Conversational AI. In *International Joint Conference on Artificial Intelligence*, pages 5844–5846, 2018.

[105] T. J. Meo, C. Kim, A. Raghavan, A. Tozzo, D. A. Salter, A. Tamrakar, and M. R. Amer. Aesop: A visual storytelling platform for conversational AI and common sense grounding. *AI Communications*, 32(1):59–76, 2019.

[106] A. H. Michaely, X. Zhang, G. Simko, C. Parada, and P. Aleksic. Keyword spotting for Google assistant using contextual speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 272–278. IEEE, 2017.

[107] T. Mikkonen and A. Taivalsaari. Web applications–spaghetti code for the 21st century. In *Sixth International Conference on Software Engineering Research, Management and Applications*, pages 319–328. IEEE, 2008.

[108] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *International Conference on Intelligent User Interfaces*, pages 397–407, 2019.

[109] M. Minsky. *Society of Mind*. Simon and Schuster, 1988.

[110] P. Mohagheghi and R. Conradi. Quality, productivity and economic benefits of software reuse: a review of industrial studies. *Empirical Software Engineering*, 12(5):471–516, 2007.

[111] S. Mohammad and P. Turney. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, 2010.

[112] C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2019.

[113] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. *Computing Research Repository*, abs/1502.00956, 2015.

[114] Muvizu. Muvizu Animation Software, 2018.

[115] C. Nicodeme. Build confidence and acceptance of AI-based decision support systems-Explainable and liable AI. In *International Conference on Human System Interaction (HSI)*, pages 20–23. IEEE, 2020.

[116] R. Nishihara, P. Moritz, S. Wang, A. Tumanov, W. Paul, J. Schleier-Smith, R. Liaw, M. Niknami, M. I. Jordan, and I. Stoica. Real-time machine learning: The missing pieces. In *Workshop on Hot Topics in Operating Systems*, pages 106–110, 2017.

[117] P. V. Ogren. Knowtator: a protégé plug-in for annotated corpus construction. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 273–275. Acl, 2006.

[118] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The building blocks of interpretability. In *Distill Publication*, 2018.

[119] K. Paranjape, M. Schinkel, R. N. Panday, J. Car, and P. Nanayakkara. Introducing artificial intelligence training in medical education. *JMIR Medical Education*, 5(2):e16048, 2019.

[120] S. Penney, J. Dodge, C. Hilderbrand, A. Anderson, L. Simpson, and M. Burnett. Toward foraging for understanding of StarCraft agents: An empirical study. In *International Conference on Intelligent User Interfaces*, pages 225–237. ACM, 2018.

[121] I. E. Perera, J. F. Allen, L. Galescu, C. M. Teng, M. H. Burstein, S. E. Friedman, D. D. McDonald, and J. M. Rye. Natural Language Dialogue for Building and Learning Models and Structures. In *Association for the Advancement of Artificial Intelligence*, 2017.

[122] W. Pieters. Explanation and trust: what to tell the user in security and AI? *Ethics and Information Technology*, 13(1):53–64, 2011.

[123] C. Ploder. Artificial Intelligence Tool Penetration in Business: Adoption, Challenges and Fears. In *Knowledge Management in Organizations*, volume 1027, page 259. Springer, 2019.

[124] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach. Manipulating and measuring model interpretability. In *CHI Conference on Human Factors in Computing Systems*, pages 1–52, 2021.

[125] D. Quick, D. Burke, and C. Kim. MUSICA: Musical Interactive Collaborative Agent. In *Conference on AI Music Creativity, AIMC*, 2021.

[126] D. Quick and C. N. Burrows. Evaluating Natural Language for Musical Operations. *International Computer Music Conference*, 2018.

[127] D. Quick and P. Hudak. Grammar-based automated music composition in Haskell. In *ACM SIGPLAN International Workshop on Functional Art, Music, Modeling, and Design*, pages 59–70, 2013.

[128] D. Quick and C. T. Morrison. Composition by Conversation. In *International Computer Music Conference*. Michigan Publishing, 2017.

[129] D. Quick and K. Thomas. A functional model of jazz improvisation. In *ACM SIGPLAN International Workshop on Functional Art, Music, Modeling, and Design*, pages 11–21, 2019.

[130] N. L. Randrup and R. O. Briggs. Evaluating the performance of collaboration engineers. In *Hawaii International Conference on System Sciences*, pages 600–609. IEEE, 2015.

[131] S. Ransbotham, D. Kiron, P. Gerbert, and M. Reeves. Reshaping business with artificial intelligence: Closing the gap between ambition and action. *MIT Sloan Management Review*, 59(1), 2017.

[132] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *Computing Research Repository*, abs/1506.02640, 2015.

[133] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In *International Conference on Machine Learning and Applications and Workshops*, volume 2, pages 241–244. IEEE, 2011.

[134] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*, 2016.

[135] R. M. Rilke and J. Burnham. Letters to a Young Poet. 1934. 1993.

[136] M. Ringgaard, R. Gupta, and F. C. N. Pereira. SLING: A framework for frame semantic parsing. *Computing Research Repository*, abs/1710.07032, 2017.

[137] K. Robertson, C. Khoo, and Y. Song. To Surveil and Predict: A Human Rights Analysis of Algorithmic Policing in Canada. 2020.

[138] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77(1):157–173, May 2008.

[139] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, 2015.

[140] R. Sanghrajka, D. Hidalgo, P. P. Chen, and M. Kapadia. LISA: Lexically Intelligent Story Assistant. *Artificial Intelligence and Interactive Digital Entertainment*, 2017.

[141] M. Schaekermann, G. Beaton, E. Sanoubari, A. Lim, K. Larson, and E. Law. Ambiguity-aware AI Assistants for Medical Data Analysis. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[142] J. Schaffer, C. Playa Vista, J. ODonovan, J. Michaelis, M. Adelphi, A. Raglin, and T. Höllerer. I can do better than your AI: Expertise and explanations. In *International Conference on Intelligent User Interfaces*, pages 240–251. ACM, 2019.

[143] I. Seeber, E. Bittner, R. O. Briggs, T. de Vreede, G.-J. De Vreede, A. Elkins, R. Maier, A. B. Merz, S. Oeste-Reiß, N. Randrup, et al. Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57(2):103174, 2020.

[144] M. Shelley and M. Butler. *Frankenstein, Or, The Modern Prometheus: The 1818 Text*. Oxford World's Classics Series. Oxford University Press, 1998.

[145] B. Shneiderman. Human values and the future of technology: A declaration of empowerment. *ACM SIGCAS Computers and Society*, 20(3):1–6, 1990.

[146] T. Shore and G. Skantze. Using Lexical Alignment and Referring Ability to Address Data Sparsity in Situated Dialog Reference Resolution. *Empirical Methods in Natural Language Processing*, 2018.

[147] L. F. Sikos. RDF-powered semantic video annotation tools with concept mapping to Linked Data for next-generation video indexing: a comprehensive review. *Multimedia Tools and Applications*, 76(12):14437–14460, 2017.

[148] R. Simon. Annotorious - Image Annotation for the Web. https://annotorious.github.io/, 2018. [Online; accessed 21-March-2018].

[149] K. Sokol and P. A. Flach. Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant. In *International Joint Conference on Artificial Intelligence*, pages 5868–5870, 2018.

[150] P. Stenetorp, S. Pyysalo, and G. Topi. Brat Rapid Annotation Tool. http://brat.nlplab.org/, 2012. [Online; accessed 21-March-2018].

[151] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition*, 2014.

[152] V. Theodorou, A. Abelló, M. Thiele, and W. Lehner. Frequent patterns in ETL workflows: An empirical approach. *Data & Knowledge Engineering*, 112:1–16, 2017.

[153] C. M. University. CMU Graphics Lab Motion Capture Database, Accessed 2018.

[154] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. In *Journal on Machine Learning Research*, volume 9, pages 2579–2605, 2008.

[155] P. Vicol, M. Tapaswi, L. Castrejon, and S. Fidler. MovieGraphs: Towards Understanding Human-Centric Situations from Videos. *Computing Research Repository*, abs/1712.06761, 2015.

[156] E. F. Villaronga, P. Kieseberg, and T. Li. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2):304–313, 2018.

[157] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*, 2014.

[158] A.-L. Vollmer, R. Read, D. Trippas, and T. Belpaeme. Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics*, 3(21):eaat7111, 2018.

[159] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[160] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently Scaling up Crowdsourced Video Annotation. *International Journal of Computer Vision*, 101(1):184–204, Jan 2013.

[161] Wagner, Johannes and Baur, Tobias and Zhang, Yue, and Valstar, Michel F. and Schuller, Björn and André, Elisabeth. Applying Cooperative Machine Learning to Speed Up the Annotation of Social Signals in Large Multi-modal Corpora. *arXiv:1802.02565*, 2018.

[162] R. Wang, D. Sun, G. Li, M. Atif, and S. Nepal. Logprov: Logging events as provenance of big data analytics pipelines with trustworthiness. In *IEEE International Conference on Big Data*, pages 1402–1411, 2016.

[163] H. J. Wilson and P. R. Daugherty. Collaborative intelligence: humans and AI are joining forces. *Harvard Business Review*, 96(4):114–123, 2018.

[164] T. Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1–191, 1972.

[165] G. N. Yannakakis, A. Liapis, and C. Alexopoulos. Mixed-initiative co-creativity. 2014.

[166] K. Yu, S. Berkovsky, R. Taib, J. Zhou, and F. Chen. Do I trust my machine teammate?: An investigation from perception to decision. In *International Conference on Intelligent User Interfaces*, pages 460–468. ACM, 2019.

[167] J. Yuen, B. Russell, C. Liu, and A. Torralba. LabelMe Video: Building a video database with human annotations. In *IEEE International Conference on Computer Vision*, pages 1451–1458, 2009.

[168] T. Zahavy, N. B. Zrihem, and S. Mannor. Graying the black box: Understanding DQNs. In *International Conference on Machine Learning*, 2016.

[169] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2014.

[170] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.

[171] J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. In *Transactions on Visualization and Computer Graphics*, volume 19, pages 2080–2089, 2013.

Chapter 9

Supplementary Materials

Questionnaire responses to AI explanations

Figures 9.1 through 9.2 provide a summary of questionnaire responses pertaining to AI explanations presented to the participant throughout the study session. Each participant was asked to evaluate whether the XAI system understands the query, and to rate their confidence and trust in the system. These responses are accompanied by explanation quality ratings collected from an external group of annotators, and then sorted and clustered by the overall level of participant satisfaction per keyword. Poorly received keywords are marked by particularly lower ratings from study participants and external annotators.

Synchronization and trust and reliance in AI explanation

Figures 9.3 through 9.6 illustrate the level of synchronization between user and AI responses as well as each user's trust and reliance in AI explanation across different experimental conditions. Icons indicate whether the user and/or the AI's answers match the ground truth for each trial, and whether the two answers overlap, indicating user-machine synchronization. Questionnaire responses pertaining to the user's experience with the XAI system are also indicated below the icons, along with the query used in each trial. Color of the query indicates the quality of generated videos based on external ratings.

Clip name	Study Rating Q1 AI understands	Study Rating Q2 Confidence in AI	Study Rating Q3 Trust in AI	External Rating Accurate Representation
express joy	4.50	4.50	4.50	4.25
salsa dance	4.75	4.00	4.00	4.00
sneaking	4.25	4.00	4.25	4.00
bear human subject	4.25	4.00	4.00	3.63
dribble	4.00	4.00	4.00	3.13
jog	4.00	4.00	4.00	2.25
punch	4.50	4.00	3.50	3.88
walk on uneven terrain	4.33	4.00	3.67	3.50
penguin human subject	4.00	4.00	3.50	2.63
walk around	4.50	3.50	3.50	4.75
walk crawl	4.00	3.75	3.75	2.25
mop	4.00	3.75	3.50	4.25
laugh	3.67	3.67	3.67	3.88
whistle walk jauntily	4.00	4.00	3.00	2.88
sit on high stool stand up	3.33	3.67	3.67	3.88
climb 3 steps	3.80	3.40	3.20	4.38
wave	3.20	3.40	3.40	2.63
wrench	3.00	3.00	4.00	2.75
clumsy walk	3.00	3.33	3.33	3.75

Figure 9.1: First half of the questionnaire response summary.

Clip name	Study Rating Q1 AI understands	Study Rating Q2 Confidence in AI	Study Rating Q3 Trust in AI	External Rating Accurate Representation
walk straight	3.00	3.33	3.33	3.50
jump	3.40	3.20	3.00	4.00
pick up	3.50	3.00	3.00	4.50
sit	3.00	3.13	3.25	3.00
yawn	3.22	3.00	2.89	1.50
walk backwards and turn	3.00	3.00	3.00	4.13
yoga	3.17	3.00	2.83	2.00
sweep floor	3.00	3.00	2.75	4.25
stretch	2.80	3.00	2.80	2.38
stretching	3.00	2.80	2.40	3.63
prairie dog human subject	3.00	2.50	2.50	3.63
sit down	2.50	2.50	3.00	3.38
stretch along	3.00	2.67	2.33	3.25
walk and turn repeated	3.00	2.00	2.50	4.63
dance	3.00	2.00	2.00	3.63
unscrew	2.50	2.00	2.00	3.13
monkey human subject	2.00	2.00	2.00	2.38
pull up	2.00	1.50	2.00	2.88
dive	1.50	1.50	1.50	1.75

Figure 9.2: Second half of the questionnaire response summary.

Figure 9.3: Participant journeys using an AI-only system.

Figure 9.4: Participant journeys using an XAI-only system.

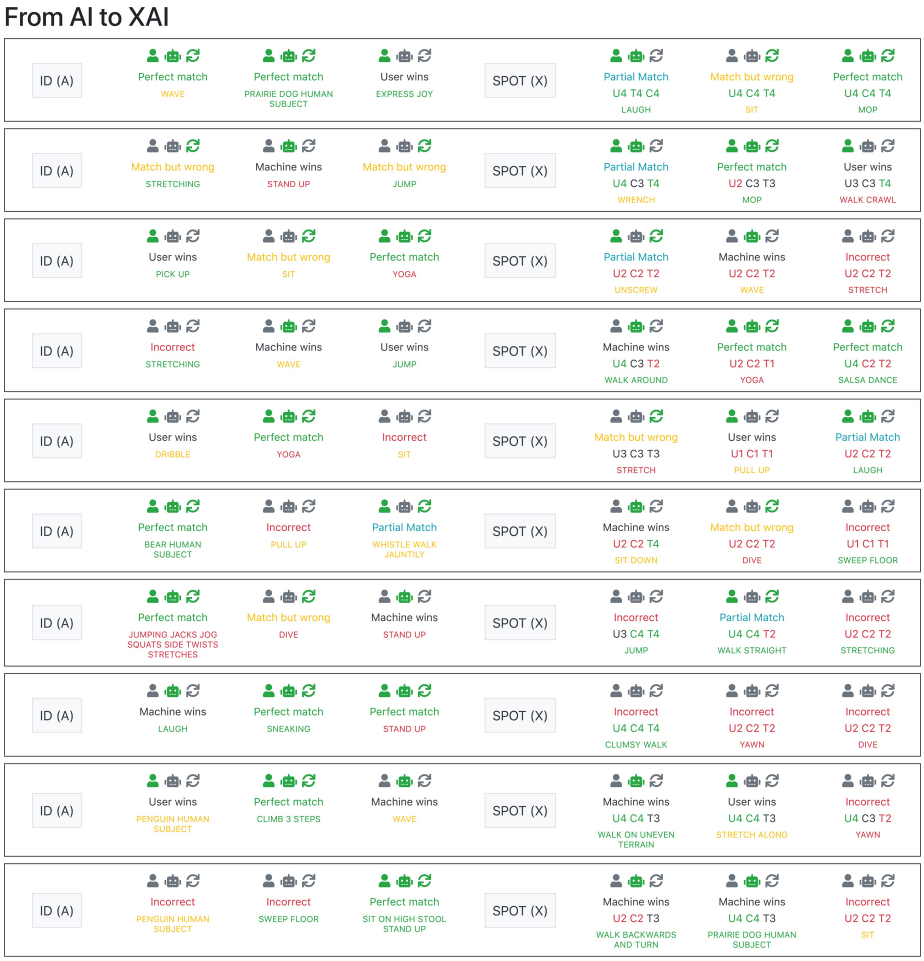

Figure 9.5: Participant journeys using an AI-only system initially then moving to an XAI system.

Figure 9.6: Participant journeys using an XAI system initially then moving to an AI-only system.