Conversational Explanations

Explainable AI through human-machine conversation

Dave Braines
CTO Emerging Technology, IBM Research UK
Industry Technical Area Leader, DAIS ITA Research Program
PhD student @
Cardiff University Crime & Security Research Institute

Full material from the original 3 hour tutorial can be downloaded from: bit.ly/conv_exp
Original Agenda

• Introductions [10]
• Explanations
  • Scene setting for Explainable AI (XAI) [20]
  • Philosophy & Social Science [20]
• Collaborative XAI research examples [10]
(Coffee break)
• Deep learning – black box explanations [20]
• The role of the user [20]
• Conversational Explanations [20]
• Visual Exploration of Deep Learning [20]
Agenda for today

• Introductions [10] [3]
• Explanations [10]
  • Scene setting for Explainable AI (XAI) [20]
  • Philosophy & Social Science [20]
• Collaborative XAI research examples [10]
(Coffee break)
• Deep learning – black box explanations [20] [5]
• The role of the user [20] [2]
• Conversational Explanations [20] [10]
• Visual Exploration of Deep Learning [20]
Introductions
Active researcher in Artificial Intelligence.

Currently focused on Machine Learning, Deep Learning and Network Motif analysis.

Published 100+ conference/journal papers.

Interested in human-machine cognitive interfaces for deep interactions between human users and machine agents.

Likes kayaking, walking and camping.
Emerging Technology, IBM Research

Delivering leading edge innovation for our clients
About us

The Crime and Security Research Institute brings together Cardiff University's significant interdisciplinary research expertise in the fields of crime and security.

The effective management of crime and security is one of the biggest challenges we face in today's world. Our response to this challenge is to conduct research on a local and global scale, combining existing academic excellence from within the Universities Police Science Institute, the Violence Research Group and the Informatics and Visual Computing Research Groups in a dynamic new initiative.

We will foster creative and innovative conceptual and methodological approaches to shape policy and practice development in relation to crime and security challenges locally, nationally and internationally; we are committed to sustaining a record of achieving real-world impact as well as addressing community concerns.
Improving Situational Understanding for Human/Machine Hybrid Teams

Dave Braines (BrainesDS@cardiff.ac.uk), 1st year PhD (part time) Supervisors: Prof. Alun Preece, Prof. Ian Taylor

Background

Machine-agent performance & human-agent confidence are increased in hybrid human-machine systems with dynamic feedback between human & machine agents.

**Human Agent Knowledge Fusion (HAKF)** is the mechanism proposed to facilitate this dynamic feedback exchange, with:

- **Explainability** providing feedback from machine agents to human users. Specifically, a description of the reasoning or processing used to reach the conclusion. This can relate to the algorithms and processes used, or can be post-hoc explanation in cases where the processing is “black box” or the algorithm details should not be shared.

- **Tellability** from the human users to the machine agents. For example to provide additional local knowledge or guidance, especially in sparse data situations which may be common in rapidly evolving situational understanding environments. This is greater than simply enhancing the training data as the situation unfolds.

All of the above is in the context of rapidly formed small coalition teams with human and machine agents, operating at the edge of the network, with limited connectivity, bandwidth and compute resources in a decision-making role.

Hypothesis

Systems with explainability will increase human-agent confidence, and systems with tellability will increase machine-agent performance. Hybrid systems with improved confidence and performance will have a measurable effect on decision making.

Narrowing the scope

Conversational explanations

Bringing together: explainability which is provided by the machine agents in the conversation, and tellability through the human agents correcting, configuring, and providing contextual information or local knowledge to improve the system.

Key 2018 Publications


Next steps

We conducted a workshop in Nov 2018 with military experts using the Design Thinking method to elicit multiple use cases for AI Explainability.

1. Complete workshop write up
2. Extend meta-model for AI Explanations
3. Refine experimental user interface
4. Plan and get approval for human trials
Distributed Analytics and Information Science
International Technology Alliance

Stanford
Purdue
Penn State
Army Research Laboratory
IBM Research
UMass Amherst
Airbus Group
Cardiff
Imperial College
IBM UK
Southampton
DSTL

Stanford
Purdue
Penn State
Army Research Laboratory
IBM Research
UMass Amherst
Airbus Group
Cardiff
Imperial College
IBM UK
Southampton
DSTL

* Government (client)
* Academic
* Industrial
Focused on rapidly formed coalitions

Running at the edge of the network
Two Technical Areas:

Dynamic, Secure Coalition Information Infrastructures

Coalition Distributed Analytics & Situational Understanding
All DAIS publications available online

bit.ly/sciencelibrary

Learning and Reasoning in Complex Coalition Information Environments: a Critical Analysis

Authors: Federico Ceruti, Mostafa Alizantot, Tianwei Xing, Dan Harborne, Jon Bakdash, Dave Braines, Supriyo Chakraborty, Lance Kaplan, Angelika Kimmig, Alun Preece, Remya Raghavendra, Mari Srivastava

Projects: BFP 55: Anticipatory Situational Understanding for Coalitions

Abstract: In this paper we provide a critical analysis with metrics that will inform guidelines for developing distributed systems for Coalition Situational Understanding (CSU).

Citations: 1

Status: Accepted

Paper Type: External Conference

Venue: FUSION 2018

Download Paper

Learning and Reasoning in Complex Coalition Information Environments: a Critical Analysis was published 1/7/2018
Explainable AI

If we want to use AI does it need to explain itself?
Defining AI

Artifacts that act like humans
Artifacts that think like humans
Artifacts that act rationally
Artifacts that think rationally

…but we’re not considering Artificial General Intelligence (AGI) today
Google computer becomes first non-human to officially qualify as car driver

New surgical robots are about to enter the operating theatre

Google Translate gets smarter with language detection, Word Lens
ARTIFICIAL INTELLIGENCE
Engineering of making Intelligent Machines and Programs

MACHINE LEARNING
Ability to learn without being explicitly programmed

DEEP LEARNING
Learning based on Deep Neural Network

https://www.edureka.co/blog/what-is-deep-learning
Fairness, Accountability, and Transparency in Machine Learning

http://www.fatml.org

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.

At the same time, there is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to “the algorithm made me do it.”
Watson (2011)

Breakthrough in “deep” question-answering via an ensemble of methods including NLP, ML, KRR ...

A key idea was that Watson tackled input questions using multiple strategies and needed a method to weigh up its certainty.
Watson (2011)

Breakthrough in “deep” question-answering via an ensemble of methods including NLP, ML, KRR ...

A key idea was that Watson tackled input questions using multiple strategies and needed a method to weigh up its certainty.
In chess, as in so many things, what computers are good at is where humans are weak, and vice versa. This gave me an idea for an experiment. What if instead of human versus machine we played as partners?

Garry Kasparov, *NY Review of Books*, 2010

“Centaur chess”
(AI + Pathologist) > Pathologist

* Error rate defined as 1 – Area under the Receiver Operator Curve
** A study pathologist, blinded to the ground truth diagnoses, independently scored all evaluation slides.

But the real surprise came when pathologists were teamed up with the Harvard team’s AI. Together, the artificial intelligence and good, ole human intelligence identified 99.5 percent of the cancerous biopsies.

© 2016 Path.AI
The “Sully” Factor

Relative strengths of computer vs. human information processing

Expertise
Knowledge
Rules
Skills
Computers
Humans

Relative strengths of computer vs. human information processing

Human+machine decision loop

- **OBSERVE**
  - Environment
  - Input data

- **ORIENT**
  - Judgement
  - Model

- **DECIDE**
  - Decision

- **ACT**
  - Action
  - Feedback data
  - Training data
Explaination points
Explanations:
Philosophy and Social Science
Key publications


Insights from the social sciences (Miller 2018)

• Humans prefer short explanations (1 or two causes)
• Contrastive explanations are best
  • Why this and not some other plausible outcome?
  • Abnormal causes are the best contrastive cases
• Explanations are selected
  • No need for a complete thorough list of causes
  • Beware: Selecting explanations can be inconsistent or contradictory
• Explanations are social interactions
  • The social context will drive the explanation content
• Explanations are truthful
  • ...and match with prior beliefs
  • ...and are generable and probable
Interpretability definitions

• “Interpretability is the degree to which a human can understand the cause of a decision” – Miller (2018)

• “Interpretability is the degree to which a human can consistently predict the models result”

• “Interpretability: the level to which an agent gains, and can make use of, both the information embedded within explanations given by the system and the information provided by the system’s transparency level.”
Interpretability considerations

- Importance/risk of a decision drives the need for interpretability
- There may be substantial additional costs for interpretability
  - As well as increased risks for privacy or adversarial attacks
- Interpretable models may be needed in cases where audit is required
  - These may be less powerful than “black box” alternatives
- Interpretation may be needed as part of the “answer”
  - In some cases the explanation qualifies the answer itself
- Decisions affecting humans or their wellbeing deserve explanations
  - GDPR has a right to explanation
- Not needed for well studies problems
- “Explanations in the wild” are becoming more commonplace
Related to interpretability

• Bias detection and mitigation
• Adversarial attacks; and defending against them
• Debugging and auditing
• Social acceptance
  • Especially of machine agents that are present in our lives
• Key considerations for interpretability:
  • Fairness
  • Privacy
  • Reliability
  • Causality
  • Trust
Interpretability methods

• Intrinsic (transparent) vs post-hoc
• Result types
  • Feature summary statistic
  • Feature summary visualization
  • Model internals
  • Data point
  • Intrinsically interpretable model
• Model specific or model agnostic
• Local or global
Interpretability techniques

• Supervised learning
  • Categorical -> classification
  • Numerical -> regression

• Interpretable models

• Model-agnostic methods
  • Surrogate models
  • LIME
  • Shapley/Shap

• Example-based explanations

• Ensemble models
Robots and programs will explain themselves
We need more intuitive interfaces to machines and programs that make heavy use of machine learning. Some examples:

- A self-driving car that reports why it stopped abruptly (“70% probability that a kid will cross the road”)
- A credit default program that explains to a bank employee why a credit application was rejected (“Applicant has too many credit cards and is employed in an unstable job”)
- A robot arm that explains why it moved the item from the conveyor belt into the trash bin (“The item has a craze at the bottom”)

These examples and more are motivating our Conversational Explanation research – a simple unified interface to support any kind of explanation...
Deep Learning

Black Box Explanations
Deep Learning - Explainability

Accuracy & Comprehensiveness

- **Input**
- **Black Box Model**
- **Prediction**
- **goldfinch**
- **Explanation**
  - e.g. SHAP

Diagram:
- Input → Black Box Model → Prediction → goldfinch → Explanation
Recap: Explanation Types and Techniques

**Explanation Types:**

- Local vs Global Explanations - The Mythos of Model Interpretability – Lipton 2016
- Transparency vs Post-Hoc - The Mythos of Model Interpretability – Lipton 2016
  (Molnar uses “intrinsic” instead of “transparent”)

**Categories:**

(with reference & expansion: Personalized explanation in machine learning – Schneider et al. 2019)

- Feature Importance (Attribution)
- Counterfactual
- Component Data
- Model Internals
- Feature Visualisation
- Explanation by Example
### Explanation Types and Techniques

**Feature Importance (Attribution)**

<table>
<thead>
<tr>
<th>LIME:</th>
<th>Shap:</th>
<th>LRP:</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>8) 05765_goldfinch.JPEG goldfinch</th>
<th>LIME</th>
<th>Shap</th>
<th>LRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>goldfinch Evidence towards predicted class shown in green</td>
<td>goldfinch Evidence towards predicted class shown in blue, evidence against shown in red.</td>
<td>goldfinch Evidence towards predicted class shown in blue, evidence against shown in red.</td>
<td></td>
</tr>
</tbody>
</table>

(Explanation Table Generated Using DAIS Interpretability Framework)
### Explanation Types and Techniques

#### Feature Importance

<table>
<thead>
<tr>
<th><strong>This is a Marsh Wren because...</strong></th>
<th><strong>This is a Marsh Wren because...</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition:</strong> this bird is brown and white in color with a skinny brown beak and brown eye rings.</td>
<td><strong>Definition:</strong> this bird is brown and white in color with a skinny brown beak and brown eye rings.</td>
</tr>
<tr>
<td><strong>Explanation:</strong> this is a small brown bird with a long tail and a <strong>white eyebrow</strong>.</td>
<td><strong>Explanation:</strong> this is a small bird with a long tail and brown and black wings.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>This is a Downy Woodpecker because...</strong></th>
<th><strong>This is a Downy Woodpecker because...</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition:</strong> this bird has a white breast black wings and a red spot on its head.</td>
<td><strong>Definition:</strong> this bird has a white breast black wings and a red spot on its head.</td>
</tr>
<tr>
<td><strong>Explanation:</strong> this is a black and white bird with a <strong>red spot</strong> on its crown.</td>
<td><strong>Explanation:</strong> this is a white bird with a black wing and a black and white striped head.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>This is a Shiny Cowbird because...</strong></th>
<th><strong>This is a Shiny Cowbird because...</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition:</strong> this bird is black with a long tail and has a very short beak.</td>
<td><strong>Definition:</strong> this bird is black with a long tail and has a very short beak.</td>
</tr>
<tr>
<td><strong>Explanation:</strong> this is a black bird with a <strong>long tail feather</strong> and a pointy black beak.</td>
<td><strong>Explanation:</strong> this is a black bird with a small black beak.</td>
</tr>
</tbody>
</table>

---

Explanation Types and Techniques

Counterfactual

Class: White Necked Raven

Class: Blue-Winged Warbler

Class: Forsters Tern

Counter-Class: American Crow

Counter-Class: Common Yellowthroat

Counter-Class: Loggerhead Shrike

This is a White Necked Raven because this is a black bird with a white nape and a large beak. This is not an American Crow because it does not have a pointy black beak.

This is a Blue Winged Warbler because this is a yellow bird with a black wing and a black pointy beak. This is not a Common Yellowthroat because it does not have a black face.

Generating Counterfactual Explanations with Natural Language – Hendricks et al. 2018
Explanation Types and Techniques

Component Data

Output To the User
Model’s Prediction: $1,600,000

Data:
- Bathrooms: 2
- Square Footage: 1140

Manipulating and Measuring Model Interpretability - Poursabzi-Sangdeh 2018
Explanation Types and Techniques

Component Data

Detecting Traffic Congestion Using a Distributed System

System Output

Prediction:
Road is Congested

Component Data:
CNN CLASSIFIER
- CNN Prediction: 0.79 Congested

Congestion Reasoner
- Congestion Rating: 0.67
---- Optical Flow: 2.3
---- Speed Limit: 30 MPH

Integrating Learning and Reasoning Services for Explainable Information Fusion – Harborne et al. 2017
Explanation Types and Techniques
Model Internals

3D visualization of a Convolution Neural Network - http://scs.ryerson.ca/~aharley/vis/fc/
Explanation Types and Techniques

Feature Visualization

Different optimization objectives show what different parts of a network are looking for.

- layer index
- x, y spatial position
- channel index
- class index

Dataset Examples show us what neurons respond to in practice.

Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.

Baseball—or stripes?
mixed4a, Unit 6

Animal faces—or snouts?
mixed4a, Unit 240

Feature Visualization - Olah, et al. 2017
Explanation Types and Techniques
Feature Visualization

Exploring Neural Networks with Activation Atlases - Carter, et al. 2019 (March 6, 2019)
Explanation Types and Techniques
Explanation by Example

Understanding Dog Vs Fish Classification Using Influence Functions

Test Image

Helpful ("influential") Images from Training Data

Understanding Black-box Predictions via Influence Functions - Koh et al. 2017
Explanation Types and Techniques
Counterfactual Explanation by Examples

Understanding Dog Vs Fish Classification Using Influence Functions

Test Image

Helpful (“influential”) Images from Training Data

Understanding Black-box Predictions via Influence Functions - Koh et al. 2017
Explanation Types and Techniques

Combinations

To understand multiple layers together, we would like each layer's factorization to be "compatible"—to have the groups of earlier layers naturally compose into the groups of later layers. This is also something we can optimize the factorization for.

The Building Blocks of Interpretability - Olah, et al. 2018
Explanation Properties

- Complexity
- Prioritization of decision information
- Visualisation of Data
- Interactivity
What makes a good explanation technique?

Desirables of Explanations

**Effectiveness:**
- Explainability (Accuracy & Comprehensiveness)
- Interpretability

**Versatility:**
- Generalizability (how many models does it work for?)
- Explanatory Power (How many questions can it answer?)

**Constraints:**
- Privacy
- Resources
- Timely
- Information Collection Effort [for personalisation]

with reference & expansion: Personalized explanation in machine learning – Schneider et al. 2019
Interpretability
Aspects of a User

- Prior Knowledge
  - Machine Learning Knowledge
  - Task Domain Knowledge
- Decision Information
- Preference
- Purpose
Experimentation Framework – Our Interface

- **Dataset Selection: Gun Wielding Image Classification**
  - **Gun Wielding Image Classification**
    - Image classification of people holding guns.
  - **Traffic Congestion Image Classification**
    - Image classification of traffic camera images.
  - **Traffic Congestion Image Classification (Resized)**
    - Resized version of the traffic congestion image classification.
  - **CIFAR-10**
    - Dataset commonly used for benchmarking Machine Learning techniques.

- **Model Selection: vgg16_imagenet**
<table>
<thead>
<tr>
<th>Model Name</th>
<th>Description</th>
<th>Performance Notes</th>
<th>Use Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv5VM</td>
<td>A keras api VGG11 CNN feature descriptor trained on Imagenet with newly trained fully connected layers.</td>
<td>Training Time: 228.53, Test Accuracy: 0.8701895</td>
<td><img src="#" alt="Use Model" /></td>
</tr>
<tr>
<td>VGG16Imagenet</td>
<td>A keras api VGG11 CNN feature descriptor trained on Imagenet with newly trained fully connected layers.</td>
<td>Training Time: 1064, Test Accuracy: 0.86</td>
<td><img src="#" alt="Use Model" /></td>
</tr>
<tr>
<td>VGG19Imagenet</td>
<td>A keras api VGG19 CNN feature descriptor trained on Imagenet with newly trained fully connected layers.</td>
<td>Training Time: 790, Test Accuracy: 0.85</td>
<td><img src="#" alt="Use Model" /></td>
</tr>
<tr>
<td>InceptionV3Imagenet</td>
<td>A keras api InceptionV3 CNN feature descriptor trained on Imagenet with newly trained fully connected layers.</td>
<td>Training Time: 508, Test Accuracy: 0.75</td>
<td><img src="#" alt="Use Model" /></td>
</tr>
</tbody>
</table>

- **Interpretability Technique: Influence Functions**
<table>
<thead>
<tr>
<th>Interpretability Technique</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIME</td>
<td>A local (example specific) decision boundary explanation of instance towards classes.</td>
</tr>
<tr>
<td>SHAP</td>
<td>An explanation by example method that finds accurate approximations of the difference in loss of a test image due to retraining the model with the exclusion of a test image.</td>
</tr>
<tr>
<td>Influence Functions</td>
<td><img src="#" alt="Influence Functions" /></td>
</tr>
<tr>
<td>LBP</td>
<td><img src="#" alt="LBP" /></td>
</tr>
</tbody>
</table>
The role of the user
“Interpretable to Whom?” framework

Argues that a machine learning system’s interpretability should be defined in relation to a specific agent or task: we should not ask if the system is interpretable, but to whom is it interpretable.

WHI workshop at ICML 2018
https://arxiv.org/abs/1806.07552
Applied to six real-world example scenarios

- Web Advertising
- Route planning on a smartphone
- Loan application
- Medical advice for clinicians
- Releasing defendants on bail
- No-go order in a military operation

...with the various roles defined in detail for each
Impact of this work

• A useful framework for assessing AI/ML system development plans and architectures

• Interest from the UK Financial Conduct Authority (FCA)
  • Invited guest lecture
  • Panel session on Ethics in AI
  • Interest in DAIS ITA research more widely

• Future plans
  • To integrate the role-based model deeper into our meta-model to support conversational explanations
  • To cross-reference against more recent work (Miller, Molnar) to standardize terminology
Conversational Explanations
Earlier Research: Conversational Interaction

• Talking to machines in natural language is ideal but hard
• Controlled Natural Language as a compromise: “easy to read, harder to write”
• Let’s bring the two together:
  – Human users write NL sentences [easy to write]
  – Machine users convert to NL [easy to process]
  – Machine users respond in CNL by default [easy to read]

there is a person named p1 that is known as ‘John Smith’ and is a high value client.
Our conversational model

- We built a model of conversations in CNL – to enable interactions that flow freely between NL and CNL.

Draws on research in agent communication languages and philosophical linguistics (speech acts).
We carried out evaluations

• Field trials
• Asset allocation
• Intelligence analysis
• Coalition planning
• Crowd-sourced intelligence
• Publication analytics
We analyzed student experiments
...and worked with practitioners
Applying conversation to explanations

• We gained key insights from this previous research
  • Conversations are social and experiential
  • They can apply in a broad set of domains
  • A single interface methodology to traverse numerous systems
  • The ability to converse across domain or system boundaries
  • Multi-modal conversations are possible

• This leads to our use of conversations for our Explainable AI research

• We hope to build a robust framework and meta-model
  • ...and carry out a series of tests with human users
Conversational Explanations

Scenario and dataset
- Real-time London CCTV imagery
- Coalition context & edge processing
- Many derivative datasets possible

Explanation-oriented architecture (XOA)
- Rapid ensemble services
- Trust and confidence

Explanation types
- Transparent, post-hoc
- Multiple modalities

Conversation and roles
- We treat explanation as a conversation
- User role and task context are key
Worked Example

Using our Explanation Oriented Architecture

- Detect or infer traffic congestion
- Congestion & explanation services and flows
- Information fusion from multi-modal data sources

Three types of congestion services:
Worked Example

Using our Explanation Oriented Architecture

- Detect or infer traffic congestion
- Congestion & explanation services and flows
- Information fusion from multi-modal data sources

Three types of congestion services:
1. Congestion Image Classifier (CIC)
Worked Example

Using our Explanation Oriented Architecture
- Detect or infer traffic congestion
- Congestion & explanation services and flows
- Information fusion from multi-modal data sources

Three types of congestion services:
1. Congestion Image Classifier (CIC)
2. Entity detector (ED)
Worked Example

Using our Explanation Oriented Architecture
• Detect or infer traffic congestion
• Congestion & explanation services and flows
• Information fusion from multi-modal data sources

Three types of congestion services:
1. Congestion Image Classifier (CIC)
2. Entity detector (ED)
3. Congestion Speed Classifier (CSC)
Worked Example

Using our Explanation Oriented Architecture

- Detect or infer traffic congestion
- Congestion & explanation services and flows
- Information fusion from multi-modal data sources

Three types of congestion services:
1. Congestion Image Classifier (CIC)
2. Entity detector (ED)
3. Congestion Speed Classifier (CSC)
Conversations for Explanation

Explanation takes the form of a conversation

- Can be in many forms
- Focus on the abstract space
- Text-chat format
- Embedded additional modalities

Conceptual model defines key concepts and relationships

- Ontology, but for machine and human processing
- Key explanation-related concepts in “core” model
Conversations for Explanation

Additional models easily integrated

- To define the context for the explanations
- Role, task, services, etc

The ad-hoc and real-time operational environment require agile context modelling.
Conversations for Explanation

*Use case 1: Fully transparent*
Conversations for Explanation

*Use case 1: Fully transparent*

Is there congestion at Checkpoint A?
Conversations for Explanation

Use case 1: Fully transparent

Is there congestion at Checkpoint A?

No, I am confident that checkpoint A is not congested.
Conversations for Explanation

Use case 1: Fully transparent

Is there congestion at Checkpoint A?

No, I am confident that checkpoint A is not congested.

Why?
Conversations for Explanation

Use case 1: Fully transparent

Is there congestion at Checkpoint A?

No, I am confident that checkpoint A is not congested.

Why?

Live CCTV shows 12 moving objects at 80% of the speed limit which indicates it is not congested.
Conversations for Explanation

*Use case 2a: Post-hoc by Saliency*
Conversations for Explanation

Use case 2a: Post-hoc by Saliency

Is there congestion at Checkpoint B?
Conversations for Explanation

Use case 2a: Post-hoc by Saliency

Is there congestion at Checkpoint B?

I am not confident, but Checkpoint B is congested.
Conversations for Explanation

*Use case 2a: Post-hoc by Saliency*

Is there congestion at Checkpoint B?

I am not confident, but Checkpoint B is congested.

Why?
Conversations for Explanation

*Use case 2a: Post-hoc by Saliency*

**Is there congestion at Checkpoint B?**

I am not confident, but Checkpoint B is congested.

**Why?**

This live CCTV image shows the situation and I believe Checkpoint B is congested (with 62% confidence).

The green areas show the parts of the image that most indicate congestion.
Conversations for Explanation

Use case 2b: Post-hoc by Example
Conversations for Explanation

Use case 2b: Post-hoc by Example

Is there congestion at Checkpoint C?
Conversations for Explanation

Use case 2b: Post-hoc by Example

Is there congestion at Checkpoint C?

I am not confident, but Checkpoint C is congested.
Conversations for Explanation

*Use case 2b: Post-hoc by Example*

Is there congestion at Checkpoint C?  

I am not confident, but Checkpoint C is congested.

Why?
Conversations for Explanation

*Use case 2b: Post-hoc by Example*

**Is there congestion at Checkpoint C?**

I am not confident, but Checkpoint C is congested.

**Why?**

I can’t show you the live CCTV image but these other images are rated as being similarly congested.
Conversations for Explanation

Use case 3: Disagreement within services
Conversations for Explanation

Use case 3: Disagreement within services

Is there congestion at Checkpoint A?
Conversations for Explanation

*Use case 3: Disagreement within services*

Is there congestion at Checkpoint A?

I cannot be confident either way, sorry.
Conversations for Explanation

Use case 3: Disagreement within services

Is there congestion at Checkpoint A?

I cannot be confident either way, sorry.

Why?
Conversations for Explanation

Use case 3: Disagreement within services

Is there congestion at Checkpoint A?

I cannot be confident either way, sorry.

Why?

Live CCTV shows 2 moving objects at 80% of the speed limit which indicates it is not congested. But the live CCTV image is classified as congested.

These outcomes are inconsistent.

The green areas show the parts of the image that most indicate congestion.
Related work

- **Insight from Social Sciences**

- **A grammar for the development of conversational explanations?**

- **Affordances – the strengths of human and machine agents**

- **Human-Computer Collaboration to drive our conversational principles**
Future plans

• Complete version 1 development of the conversational meta-model
• Build the experimental conversational explanation capability
  • Aligned against the conversational meta-model
• Choose a domain of interest for experimentation
• Design a user-focused experiment
  • Conversational Explanations
  • Measure some impact across multiple groups to test the effectiveness of conversational explanation
Thank you for listening!

Conversational Explanations
Explainable AI through human-machine conversation

Dave Braines
dave_braines@uk.ibm.com