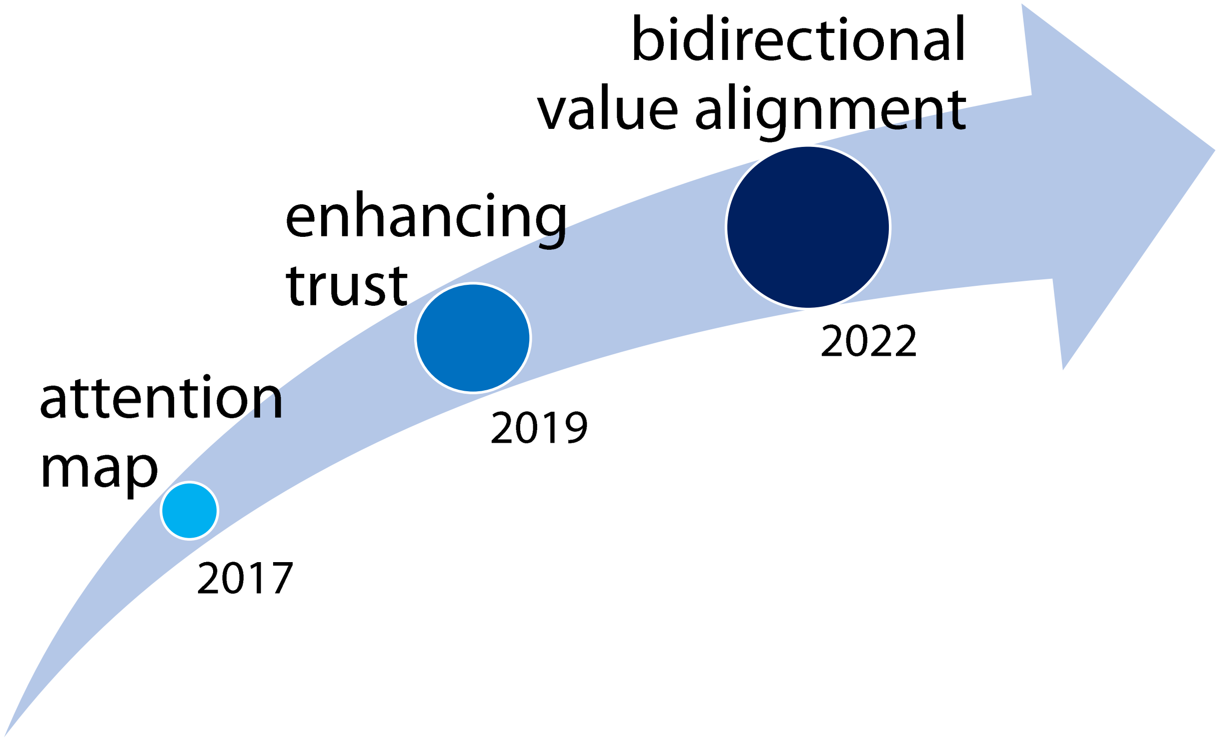
22.07.27@CASIA

# The Journey of XAI





Dr. Yixin Zhu
Peking University

yixin.zhu@pku.edu.cn
https://yzhu.io


# Explainable Artificial Intelligence (XAI)

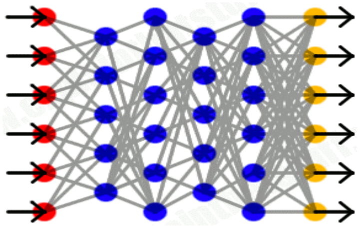
EXPLAINABLE ARTIFICIAL INTELLIGENCE

David Gunning

DARPA/I2O

# The Need for Explainable AI

AI System



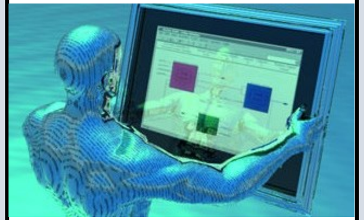
- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

**Transportation**

**Finance**

**Security**

**Legal**



**Medicine**

**Military**

User

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- The current generation of AI systems offer tremendous benefits, but their effectiveness will be limited by the machine’s inability to explain its decisions and actions to users.
- Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners.

# XAI In the News


**MIT Technology Review**

**The Dark Secret at the Heart of AI**
Will Knight
April 11, 2017

THE WALL STREET JOURNAL. WSJ

**Inside DARPA's Push to Make Artificial Intelligence Explain Itself**
Sara Castellanos and Steven Norton
August 10, 2017

The New York Times Magazine

**Can A.I. Be Taught to Explain Itself?**
Cliff Kuang
November 21, 2017

Intelligent Machines Are Asked to Explain How Their Minds Work
Richard Waters
July 11, 2017

FT FINANCIAL TIMES

The Register

You better explain yourself, mister: DARPA's mission to make an accountable AI
Dan Robinson
September 29, 2017

ExecutiveBiz

Charles River Analytics-Led Team Gets DARPA Contract to Support Artificial Intelligence Program
Ramona Adams
June 13, 2017

Entrepreneur

Elon Musk and Mark Zuckerberg Are Arguing About AI -- But They're Both Missing the Point
Artur Kiulian
July 28, 2017

Team investigates artificial intelligence, machine learning in DARPA project
Lisa Daigle
June 14, 2017

Military EMBEDDED SYSTEMS

FAST COMPANY

Why The Military And Corporate America Want To Make AI Explain Itself
Steven Melendez
June 22, 2017

NOVA NEXT

Ghosts in the Machine
Christina Couch
October 25, 2017

Jane's

DARPA's XAI seeks explanations from autonomous systems
Geoff Fein
November 16, 2017

COMPUTERWORLD

**Oracle quietly researching 'Explainable AI'**
George Nott
May 5, 2017

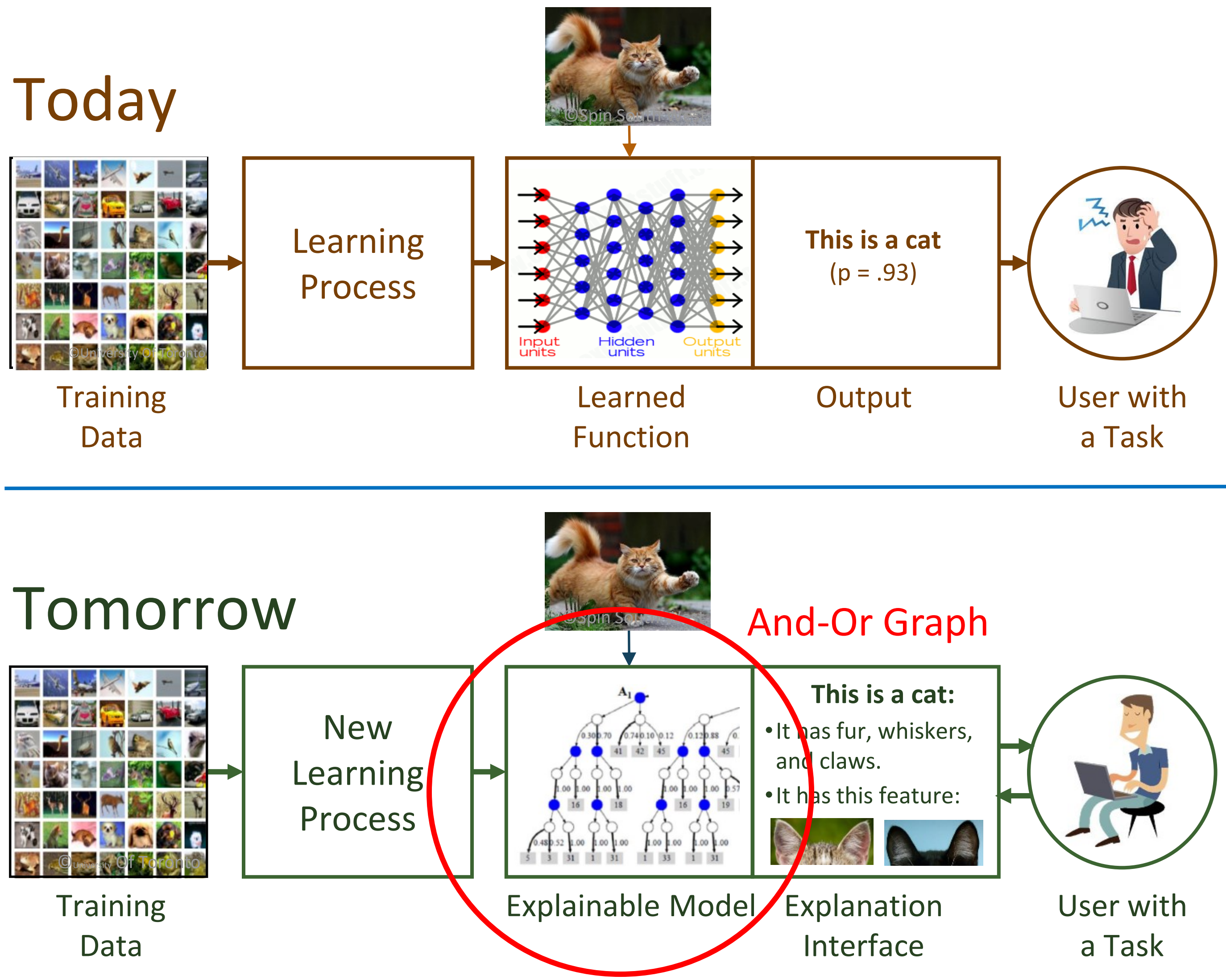
SCIENTIFIC AMERICAN.

Demystifying the Black Box That Is AI
Ariel Bleicher
August 9, 2017

How AI detectives are cracking open the black box of deep learning
Paul Voosen
July 6, 2017

Science AAAS

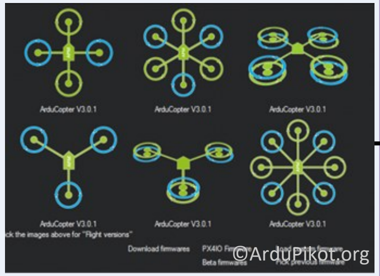
# What Are We Trying To Do?




- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

# Challenge Problems

| | Learn a model | Explain decisions | Use the explanation |
|---|---|---|---|
| **Data Analytics** Classification Learning Task |   Multimedia Data | Explainable Model → Explanation Interface |   |
| | Classifies items of interest in large data set | Explains why/why not for recommended items | Analyst decides which items to report, pursue |
| **Autonomy** Reinforcement Learning Task |   ArduPilot & SITL Simulation | Explainable Model → Explanation Interface |   |
| | Learns decision policies for simulated missions | Explains behavior in an after-action review | Operator decides which future tasks to delegate |

Data Analytics: Recommend → / ← Explanation

Autonomy: Actions → / ← Explanation

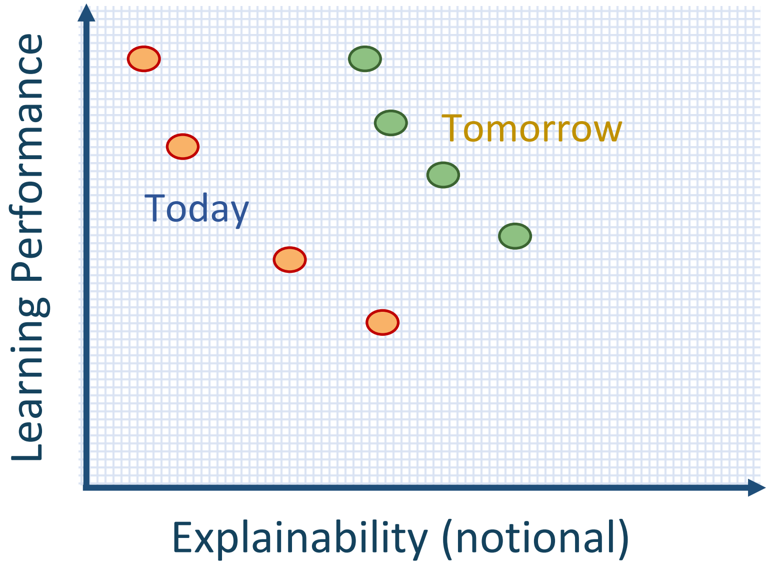
An analyst is looking for items of interest in massive multimedia data sets

An operator is directing autonomous systems to accomplish a series of missions

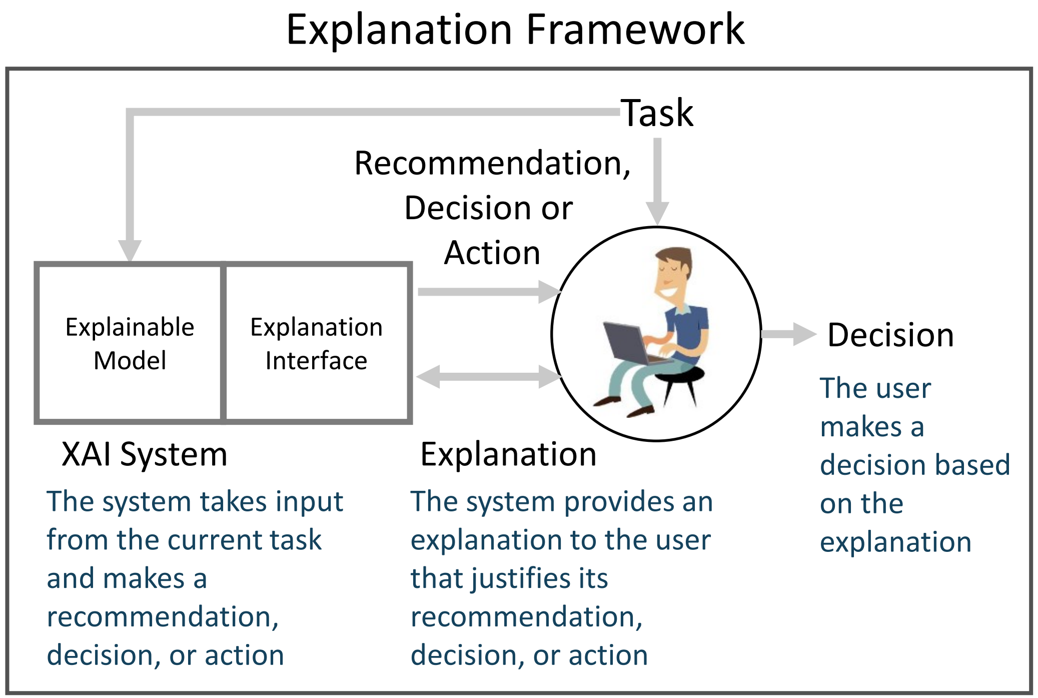
# Goal: Performance and Explainability

- XAI will create a suite of machine learning techniques that
  - Produce more explainable models, while maintaining a high level of learning performance (e.g., prediction accuracy)
  - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners

Performance vs. Explainability



# Measuring Explanation Effectiveness




| Measure of Explanation Effectiveness |
|---|
| User Satisfaction |
| • Clarity of the explanation (user rating)<br>• Utility of the explanation (user rating) |
| Mental Model |
| • Understanding individual decisions<br>• Understanding the overall model<br>• Strength/weakness assessment<br>• 'What will it do' prediction<br>• 'How do I intervene' prediction |
| Task Performance |
| • Does the explanation improve the user's decision, task performance?<br>• Artificial decision tasks introduced to diagnose the user's understanding |
| Trust Assessment |
| • Appropriate future use and trust |
| Correctability (Extra Credit) |
| • Identifying errors<br>• Correcting errors<br>• Continuous training |

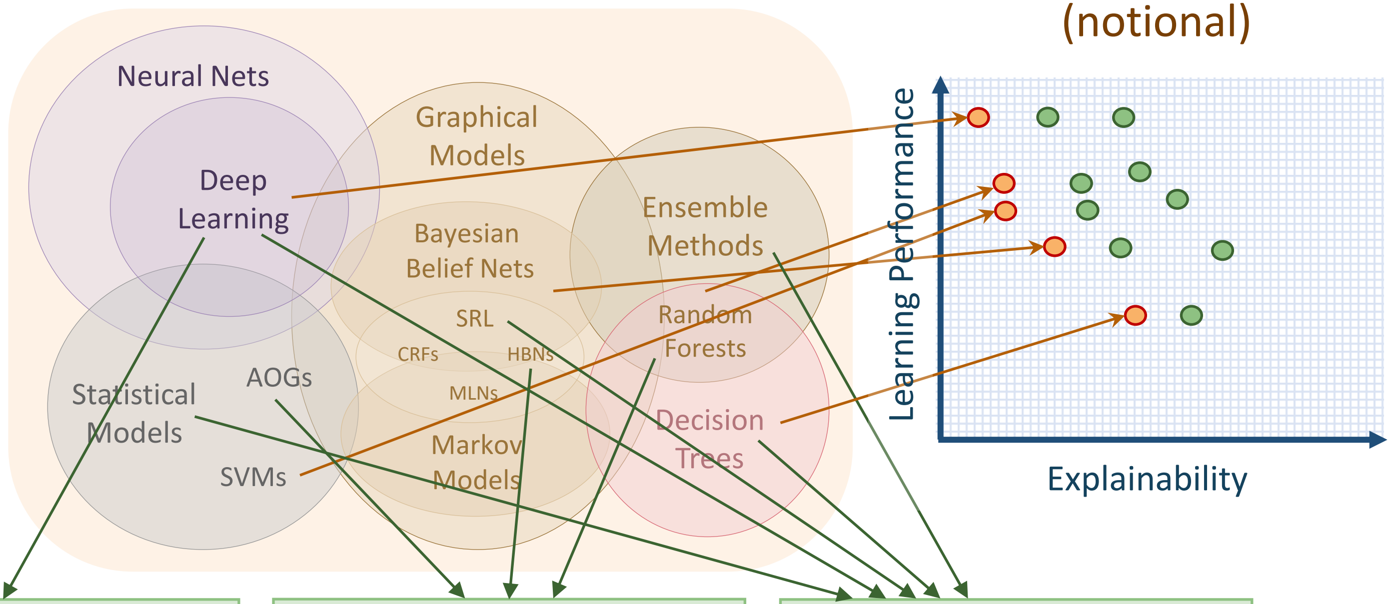
# Performance vs. Explainability

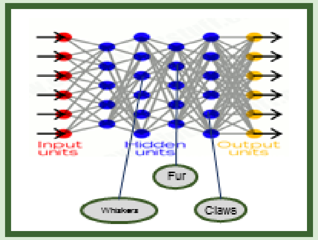
## New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

## Learning Techniques (today)

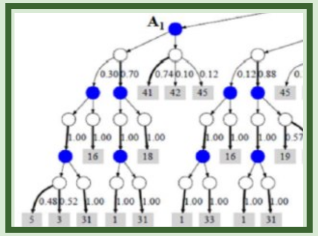
## Explainability (notional)



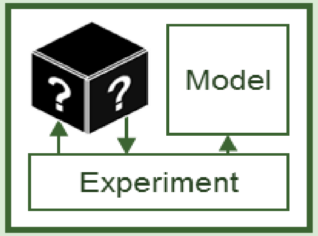

### Deep Explanation

Modified deep learning techniques to learn explainable features



### Interpretable Models

Techniques to learn more structured, interpretable, causal models

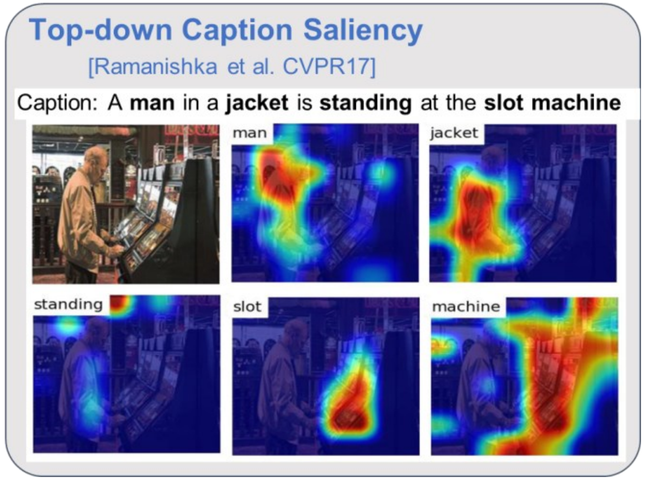



### Model Induction

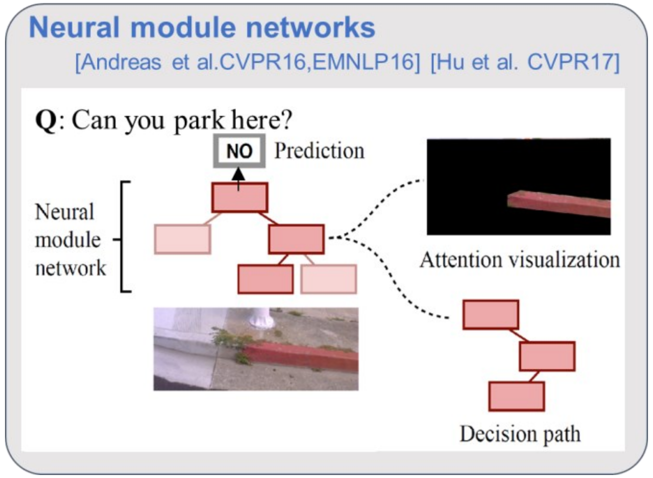
Techniques to infer an explainable model from any model as a black box

# Approaches to Deep Explanation
## (Berkeley, SRI, BBN, OSU, CRA, PARC)

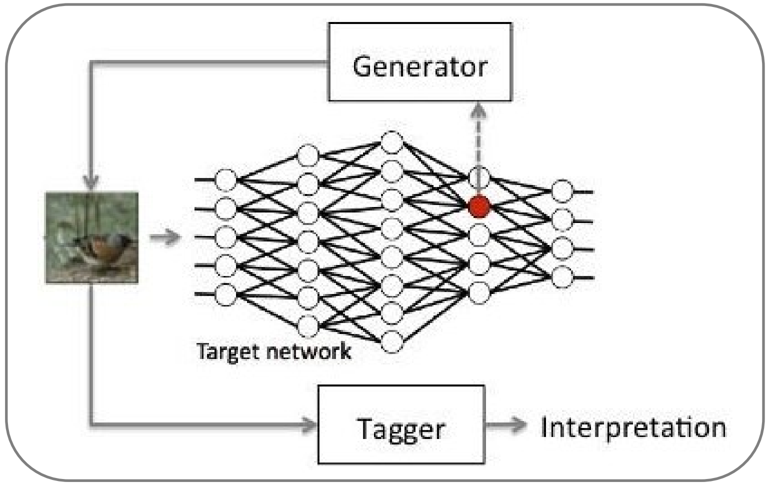
### Attention Mechanisms



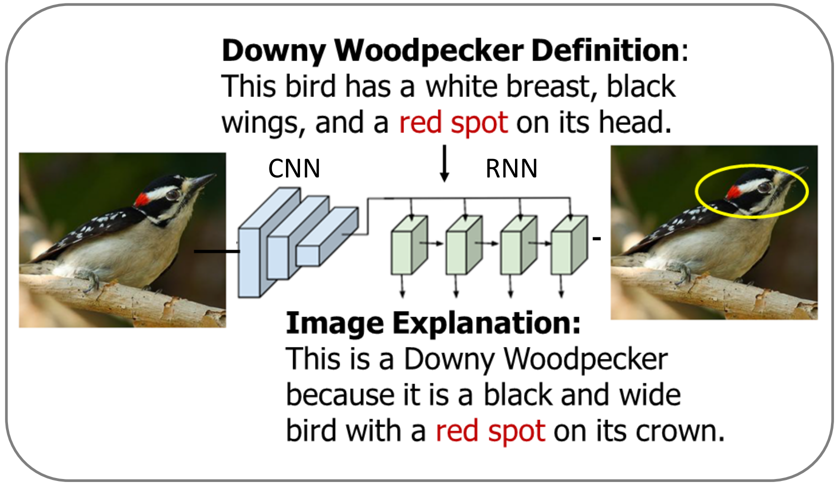

### Modular Networks




### Feature Identification




### Learn to Explain



# Deeply Explainable Artificial Intelligence

## Explainable Model

### Deep Learning

- Explain *implicit* (latent) nodes by training additional DL models
- Explain *explicit* nodes thru Neural Module Networks (NMNs).

## Explanation Interface

### Reflexive & Rational

- Reflexive explanations (that arise directly from the model)
- Rational explanations (that come from reasoning about user's beliefs)

## Challenge Problem

### Autonomy

- ArduPilot and OpenAI Gym Simulations

### Data Analytics

- Visual QA and Multimedia Event QA

- **PI**: Trevor Darrell (Berkeley)

- Pieter Abbeel (Berkeley)
- Tom Griffiths (Berkeley)
- Kate Saenko (BU)
- Zeynep Akata (U. Amsterdam)
- Dan Klein (Berkeley)
- John Canny (Berkeley)
- Anca Dragan (Berkeley)
- Anthony Hoogs (Kitware)

# DARE: Deep Attention-based Representations for Explanation

## Explainable Model

### Deep Learning

- Multiple deep learning techniques:
  - Attention-based mechanisms
  - Compositional NMNs
  - GANs

## Explanation Interface

### Show-and-Tell Explanations

- DNN visualization
- Query evidence that explains DNN decisions
- Generate natural language justifications

## Challenge Problem

### Data Analytics

- Visual Question Answering (VQA) using Visual Gnome, Flickr30
- MovieQA

- **PIs**: Giedrius Burachas (SRI), Mohamed Amer (SRI)

- Shalini Ghosh (SRI)
- Avi Ziskind (SRI)
- Michael Wessel (SRI)
- Richard R. Zemel (U. Toronto)
- Sanja Fidler (U. Toronto)
- David Duvenaud (U. Toronto)
- Graham Taylor (U. Guelph)
- Jürgen Schulze (UCSD)

# EQUAS: Explainable QUestion Answering System

## Explainable Model

### Deep Learning

- Semantic labelling of DNN neurons
- DNN audit trail construction
- Gradient-weighted Class Activation Mapping

## Explanation Interface

### Argumentation Theory

- Comprehensive strategy based on argumentation theory
- NL generation
- DNN visualization

## Challenge Problem

### Data Analytics

- Visual Question Answering (VQA), beginning with images and progressing to video

- **PI**: William Ferguson (Raytheon BBN)

- Antonio Torralba (MIT)
- Ray Mooney (UT Austin)
- Devi Parikh (GA Tech)
- Dhruv Batra (GA Tech)

# Naturalistic Decision Making Foundations of Explainable AI

## Literature Review

### Naturalistic Theory

- Extensive review of relevant psychological theories
- Extend the theory of Naturalistic Decision Making to cover explanation

## Computational Model

### Bayesian Framework

- Represent reductionist mental models that humans develop as part of the explanatory process
- Including mental simulation

## Model Validation

### Experiments

- Conduct interactive assessment and formal human experiments
- Validate the model
- Develop metrics of explanation effectiveness

- **PI:** Robert R. Hoffman (IHMC)

- Gary Klein (MacroCognition)
- Shane T. Mueller (Michigan Tech)
- William J. Clancey (IHMC)
- COL Timothy M. Cullen (SAASS)
- Jordan Litman (IHMC Psychometrician)
- Simon Attfield (Middlesex University-London)
- Peter Pirolli (IHMC)

# Tractable Probabilistic Logic Models: A New, Deep Explainable Representation

## Explainable Model

### Probabilistic Logic

- Tractable Probabilistic Logic Models (TPLMs) – an important class of (non-deep learning) interpretable models

## Explanation Interface

### Probabilistic Decision Diagrams

- Enables users to explore and correct the underlying model as well as add background knowledge

## Challenge Problem

### Data Analytics

- Infer activities in multimodal data (video and text)
- Using the Wetlab (biology) and TACoS (cooking) datasets


- **PI**: Vibhav Gogate (UT Dallas)

- Adnan Darwiche (UCLA)
- Guy Van Den Broeck (UCLA)
- Nicholas Ruozzi (UT Dallas)
- Eric Ragan (Texas A&M)
- Parag Singla (IIT-Delhi)

# XRL: Explainable Reinforcement Learning for AI Autonomy

## Explainable Model

### XRL Models

- Create a new scientific discipline for Explainable Reinforcement Learning with work on new algorithms and representations

## Explanation Interface

### XRL Interaction

- Interactive explanations of dynamic systems
- Human-machine interaction to improve performance

## Challenge Problem

### Autonomy

- Open AI Gym
- Autonomy in the electrical grid
- Mobile service robots
- Self-improving educational software


- **PI**: Geoff Gordon (CMU)

- Zico Kolter (CMU)
- Pradeep Ravikumar (CMU)
- Manuela Veloso (CMU)
- Emma Brunskill (Stanford)

# Transforming Deep Learning to Harness the Interpretability of Shallow Models: An Interactive End-to-End System

## Explainable Model

### Mimic Learning

- Develop a mimic learning framework that combines deep learning models for prediction and shallow models for explanations

## Explanation Interface

### Interactive Visualization

- Interactive visualization over multiple views, using heat maps & topic modeling clusters to show predictive features

## Challenge Problem

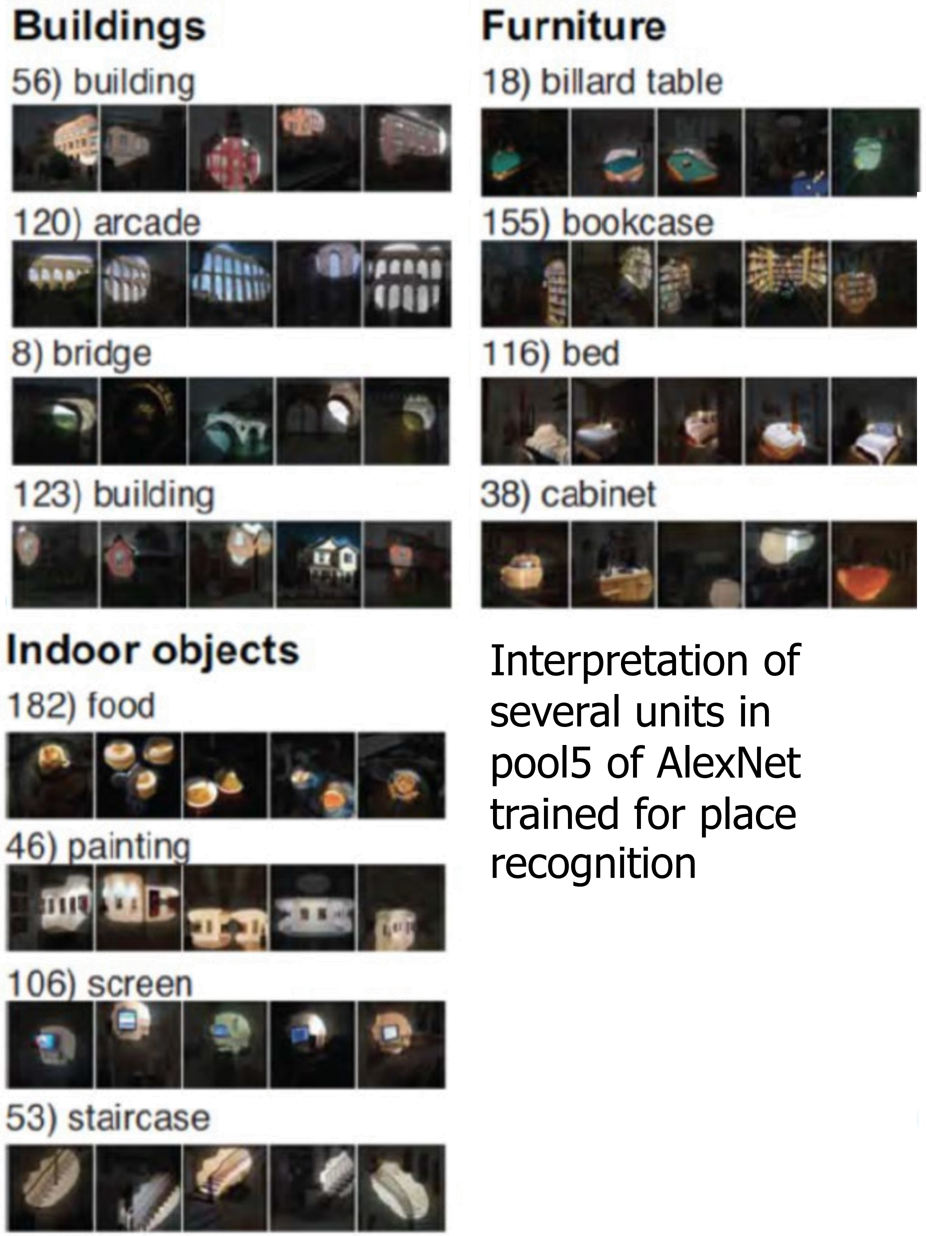
### Data Analytics

- Multiple tasks using data from Twitter, Facebook, ImageNet, UCI, NIST and Kaggle
- Metrics for explanation effectiveness


- **PI**: Xia Hu (Texas A&M)

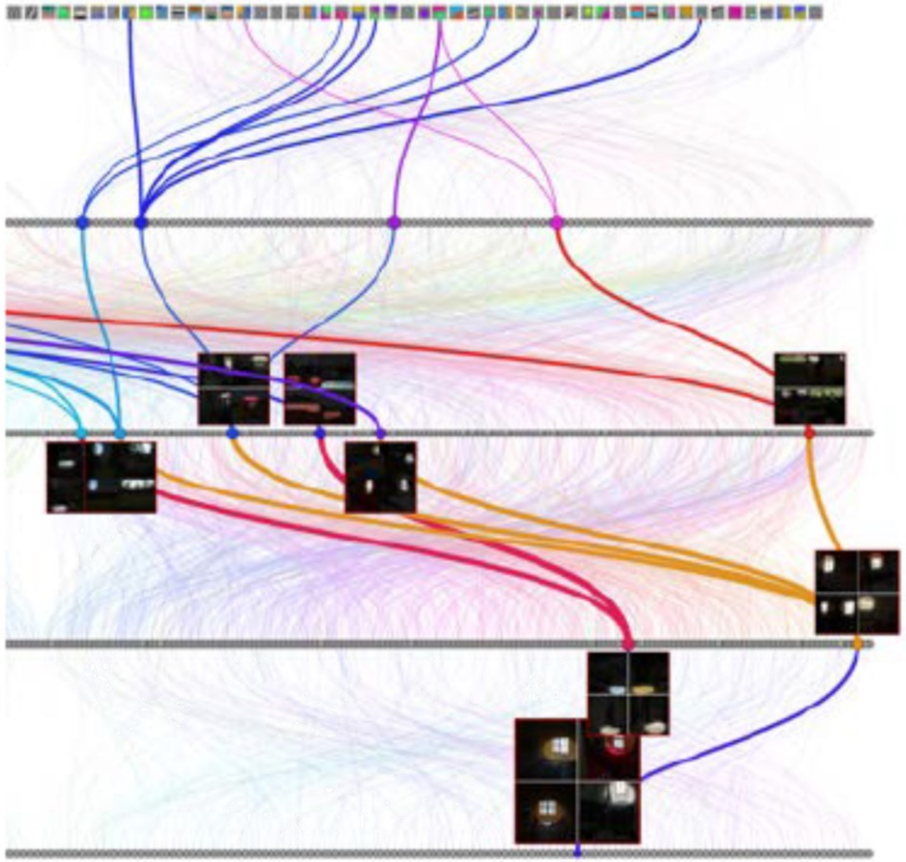
- Shuiwang Ji (Wash. State)
- Eric Ragan (Texas A&M)

# Network Dissection Quantifying Interpretability of Deep Representations (MIT)

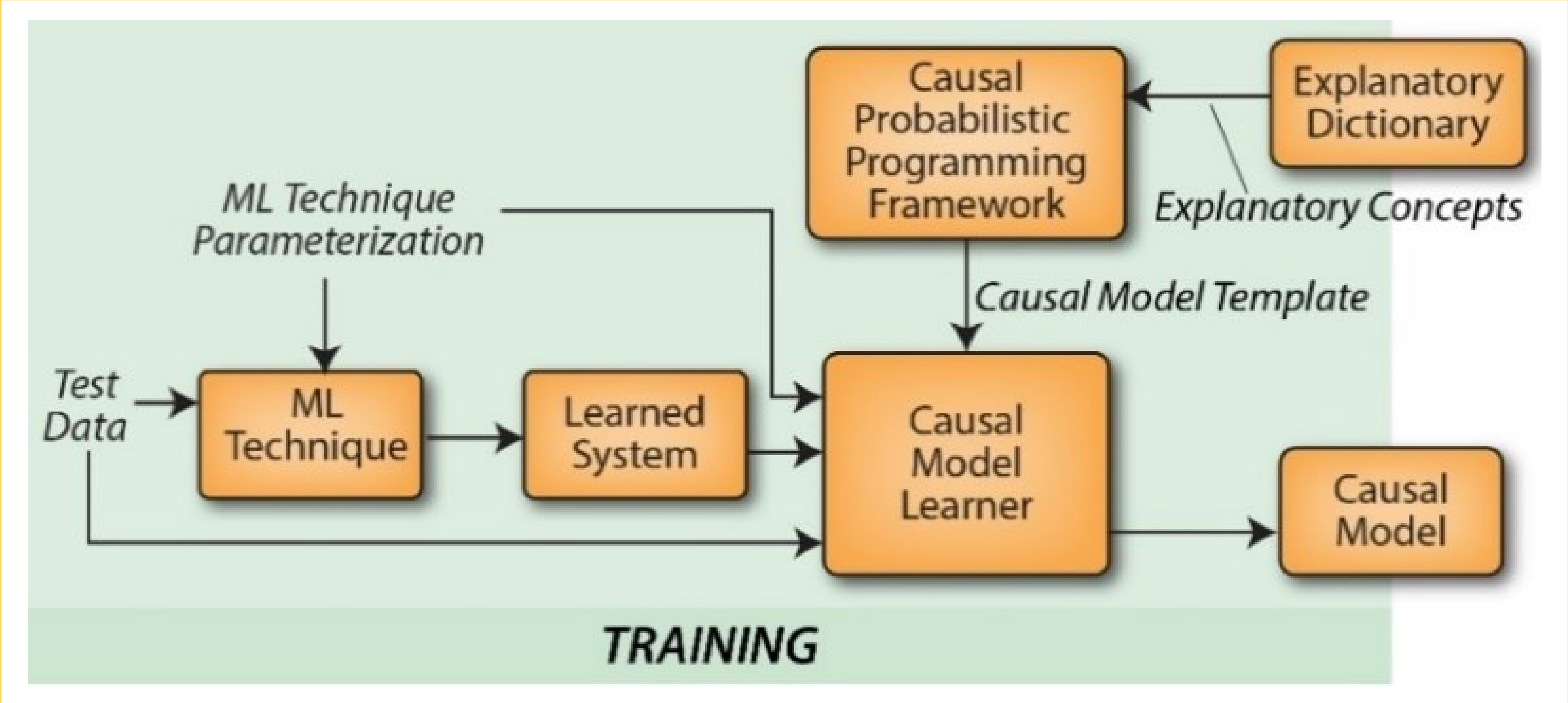



Interpretation of several units in pool5 of AlexNet trained for place recognition

Audit trail: for a particular output unit, the drawing shows the most strongly activated path



# Causal Model Induction (CRA)




<u>Causal Model Induction</u>: Experiment with the learned model (as a grey box) to learn an explainable, causal, probabilistic programming model

# CAMEL: Causal Models to Explain Learning

## Explainable Model

### Model Induction
### Causal Models

- Experiment with the learned model (as a grey box) to learn an explainable, causal, probabilistic programming model

## Explanation Interface

### Narrative Generation

- Interactive visualization based on the generation of temporal, spatial narratives from the causal, probabilistic models

## Challenge Problem

### Autonomy

- Minecraft, Starcraft

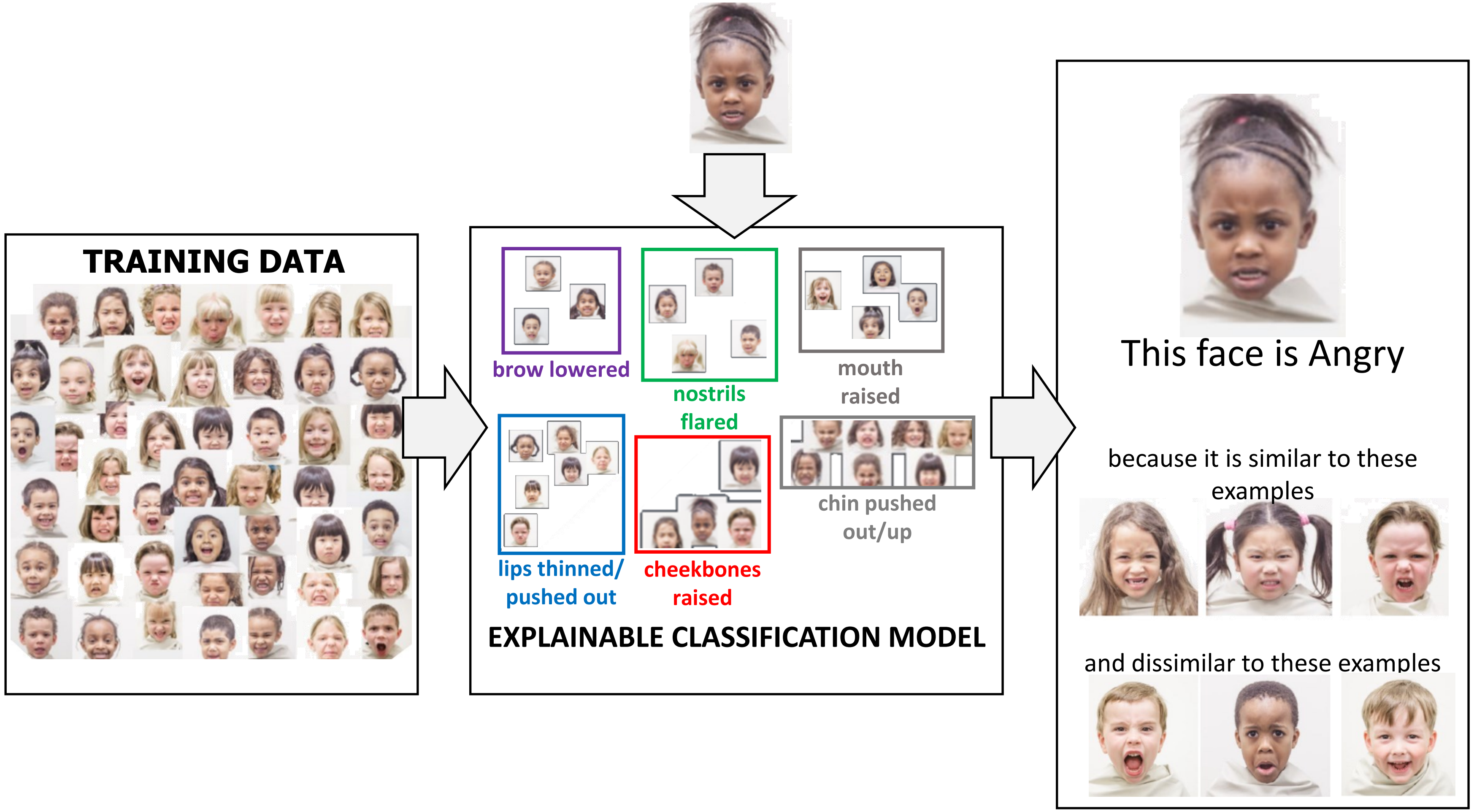
### Data Analytics

- Pedestrian Detection (INRIA), Activity Recognition (ActivityNet)

- **PI**: Brian Ruttenberg (CRA)

- Avi Pfeffer (CRA)
- David Jensen (U. Mass)
- Michael Littman (Brown)
- James Niehaus (CRA)
- Emilie Roth (Roth Cognitive Engineering
- Joe Gorman(CRA)
- James Tittle (CRA)

# Explanation by Selection of Teaching Examples (Rutgers)




**BAYESIAN TEACHING** for optimal selection of examples for machine explanation

# Model Explanation by Optimal Selection of Teaching Examples

## Explainable Model

### Model Induction

- Select the optimal training examples to explain model decisions based on Bayesian Teaching

## Explanation Interface

### Bayesian Teaching

- Example-based explanation of:
    - the full model
    - user-selected sub-structure
    - user submitted examples

## Challenge Problem

### Data Analytics

- Movie descriptions
- Image processing
- Caption data
- Movie events
- Human motion events

- **PI**: Patrick Shafto (Rutgers)

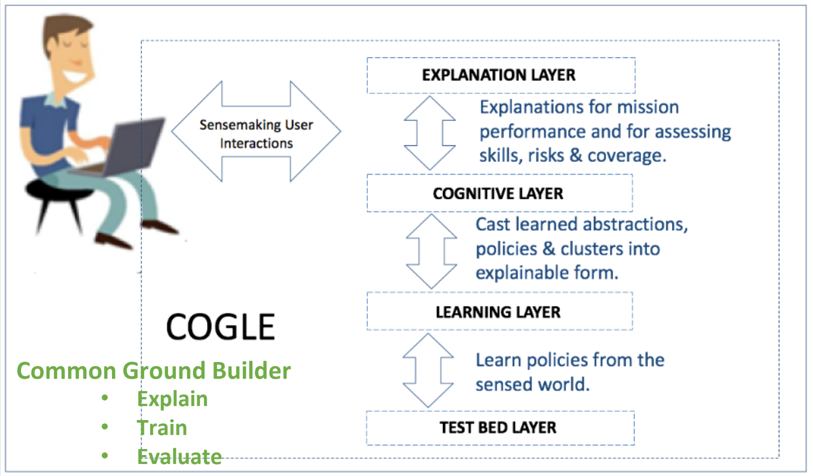
- Scott Cheng-Hsin Yang (Rutgers)

# Autonomy (PARC, OSU)

EXPLAINABLE ARTIFICIAL INTELLIGENCE

## Common Ground Learning and Explanation (COGLE)

An interactive sensemaking system to explain the learned performance capabilities of a UAS flying in an ArduPilot simulation testbed





Series 1. **Primitives:** Navigating with Constraints and Lookahead ....... 7
Lesson 1.1: **Taking off** ....... 7
Lesson 1.2: **Taking off and Landing** ....... 9
Lesson 1.3: **Reconnaissance Over a Point (3 Months)** ....... 11
Lesson 1.4: ***Looking Ahead* to Avoid Crashing into Mountains** ....... 13
Lesson 1.5: **Choosing a Safe Descent Approach for Landing** ....... 15
Lesson 1.6: **Provisioning a Hiker (6 months)** ....... 17

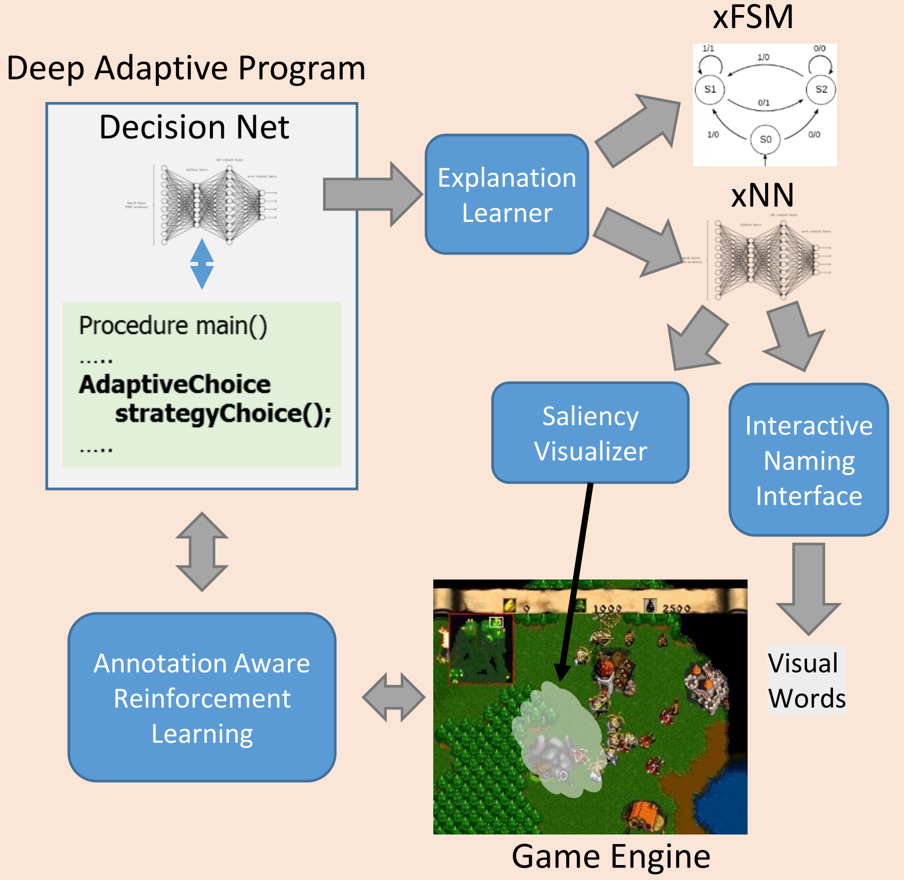
Series 2. **Behaviors:** Managing Competing Goals and Foraging ....... 19
Lesson 2.1: **Provisioning a Hiker in a Box Canyon (opt)** ....... 19
Lesson 2.2: **Taking an Inventory of a Region and Refueling (opt)** ....... 22
Lesson 2.3: **Foraging Around a Point for a Hiker (opt)** ....... 24
Lesson 2.4: **Foraging Around a Point with an Interfering Obstacle** ....... 26

Series 3. **Missions:** Harder Missions and Heavy Testing ....... 28
Lesson 3.1: **Double Hiker Jeopardy (9 months)** ....... 28
Lesson 3.2: **Bear on the Runway** ....... 30
Lesson 3.3: **Auto-Generated Missions with Testing (12 months)** ....... 32


**Robotics Curriculum**

## Explanation-Informed Acceptance Testing of Deep Adaptive Programs (xACT)

Tools for explaining deep adaptive programs and discovering best principles for designing explanation user interfaces



# COGLE: Common Ground Learning and Explanation

## Explainable Model

### Cognitive Model

- 3-layer architecture:
  - Learning Layer (DNNs)
  - Cognitive Layer (ACT-R Cog. Model)
  - Explanation Layer (HCI)

## Explanation Interface

### Interactive Training

- Interactive visualization of states, actions, policies & values
- Includes a module for test pilots to refine and train the system

## Challenge Problem

### Autonomy

- ArduPilot simulation environment
- *Value of Explanation* (VoE) framework for measuring explanation effectiveness

- **PI**: Mark Stefik (PARC)

- Sricharan Kumar (PARC)
- Honglak Lee (U. Mich.)
- Subramanian Ramamoorthy (U. Edinburgh)

- Christian Lebiere (CMU)
- John Anderson (CMU)
- Robert Thomson (USMA)

- Michael Youngblood (PARC)

# xACT: Explanation-Informed Acceptance Testing of Deep Adaptive Programs

| Explainable Model | Explanation Interface | Challenge Problem |
|---|---|---|
| **Adaptive Programs**<br>• Explainable Deep Adaptive Programs (xDAPs) – a new combination of Adaptive Programs, Deep Learning and explainability | **Acceptance Testing**<br>• Provides a visual & NL explanation interface for acceptance testing by test pilots based on Information Foraging Theory | **Autonomy**<br>• Real-Time Strategy Games based on custom designed game engine designed to support explanation<br>• Possible use of Starcraft |

- **PI**: Alan Fern (OSU)

- Tom Dietterich (OSU)
- Fuxin Li (OSU)
- Prasad Tadepalli (OSU)
- Weng-Keen Wong (OSU)
- Margaret Burnett (OSU)
- Martin Erwig (OSU)
- Liang Huang (OSU)

# UCLA/OSU/Michigan State



## Learning and Communicating Explainable Representations for Analytics and Autonomy

### Explainable Model

**Pattern Theory+**

- Integrated representation across an entropy spectrum:
  - Deep Neural Nets
  - Stochastic And-Or-Graphs (AOG)
  - Predicate Calculus

### Explanation Interface

**3-Level Explanation**

- Integrate 3 levels of explanation:
  - Concept compositions
  - Causal and counterfactual reasoning
  - Utility explanations

### Challenge Problem

**Autonomy**

- Humanoid robot behavior and VR simulation platform

**Data Analytics**

- Understanding complex multimedia events

- **PI**: Song-Chun Zhu (UCLA)

- Ying Nian Wu (UCLA)
- Sinisa Todorovic (OSU)
- Joyce Chai (Michigan State)

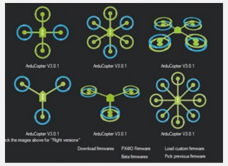
# XAI Program Structure

DARPA

XAI EXPLAINABLE ARTIFICIAL INTELLIGENCE

**Challenge Problem Areas**


**Data Analytics**
Multimedia Data

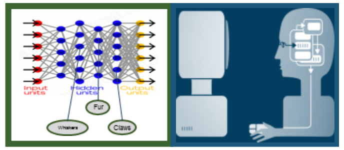


**Autonomy**
ArduPilot &
SITL Simulation

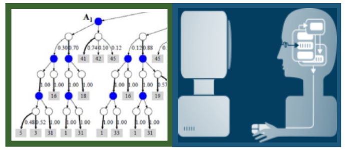
**TA 1:**
Explainable Learners

Teams that provide prototype systems with both components:

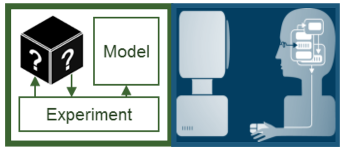
- Explainable Model
- Explanation Interface



**Deep Learning Teams**

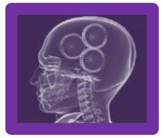


**Interpretable Model Teams**




**Model Induction Teams**

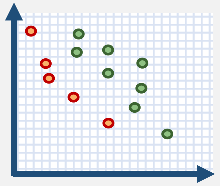
**TA 2:**
Psychological Model of Explanation



- Psych. Theory of Explanation
- Computational Model
- Consulting

**Evaluation Framework**

Learning Performance



Explanation Effectiveness

Explanation Measures

- User Satisfaction
- Mental Model
- Task Performance
- Trust Assessment
- Correctability

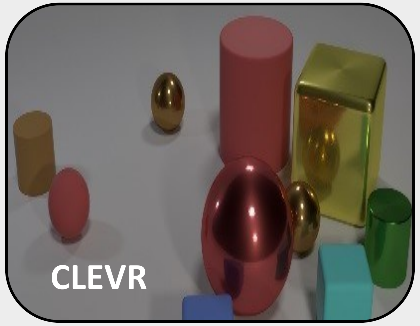
**Evaluator**
**Naval Research Lab**

- TA1: Explainable Learners
  - Multiple TA1 teams will develop prototype explainable learning systems that include both an explainable model and an explanation interface
- TA2: Psychological Model of Explanation
  - At least one TA2 team will summarize current psychological theories of explanation and develop a computational model of explanation from those theories

# Challenge Problem Candidates

## Analytics

### Visual Question Answering

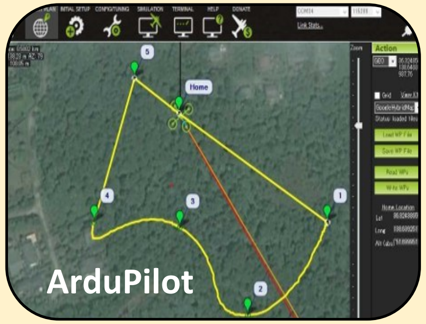

### Activity Recognition


## Autonomy

### Strategy Games



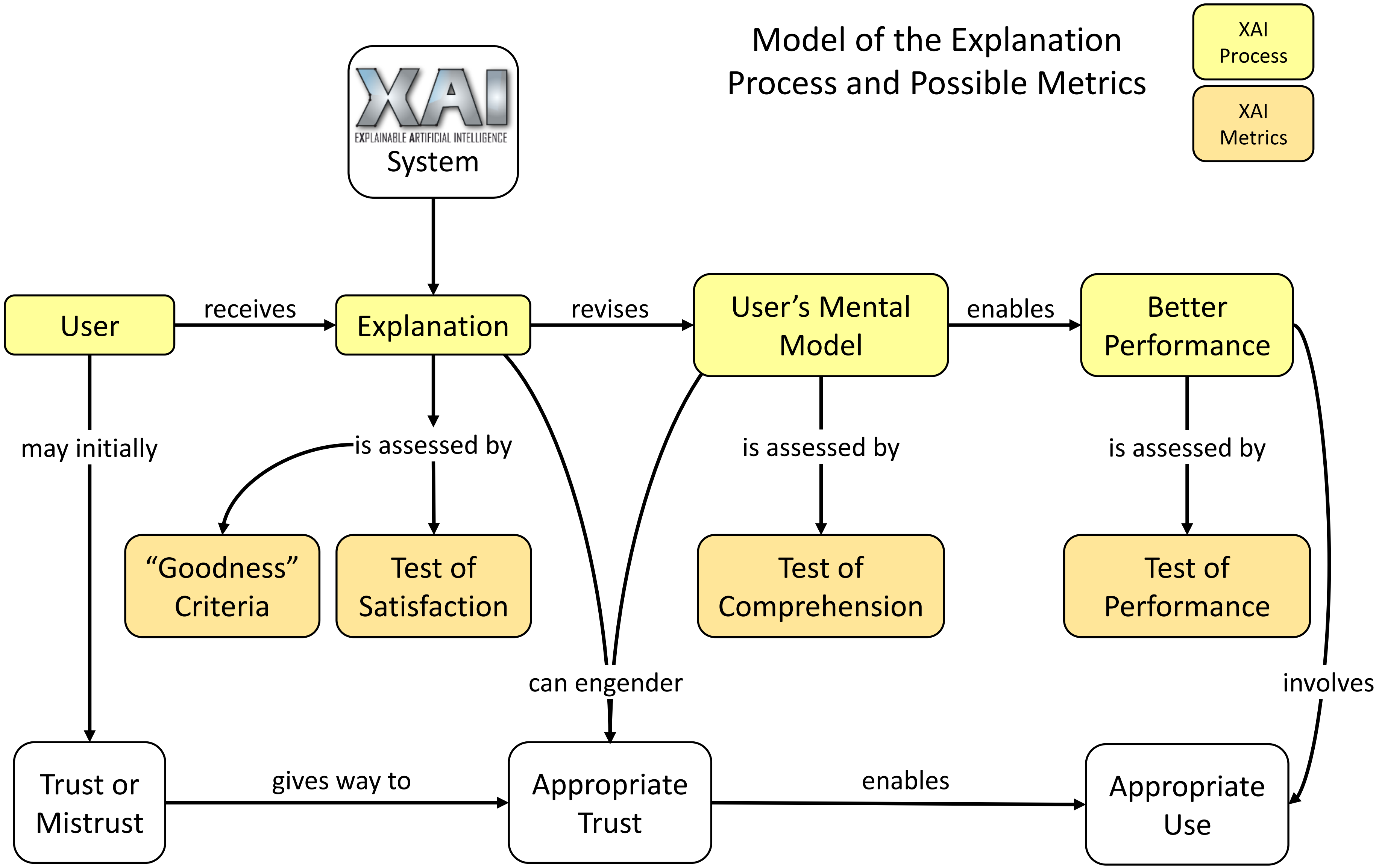
### Vehicle Control



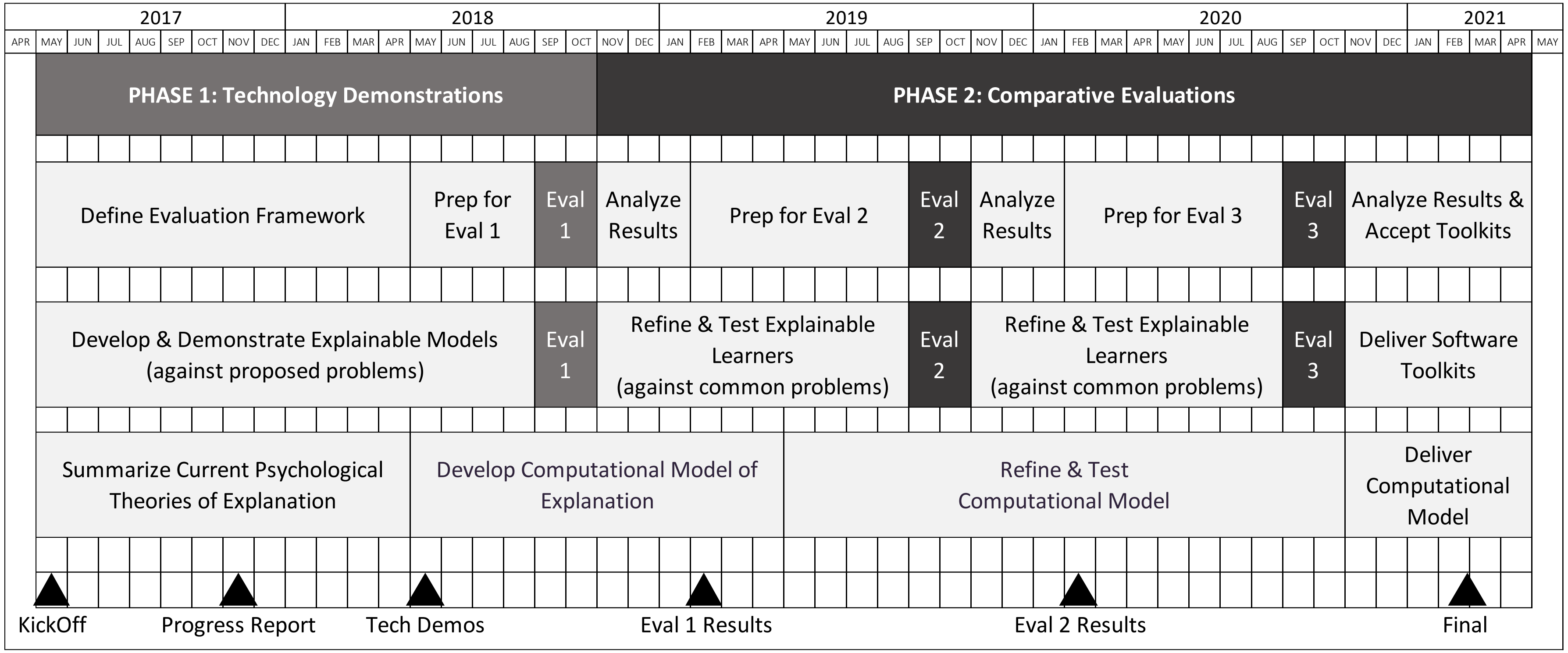

DARPA

# IHMC/MacroCognition/Michigan Tech
## Psychological Models of Explanation

XAI EXPLAINABLE ARTIFICIAL INTELLIGENCE



# Schedule and Milestones



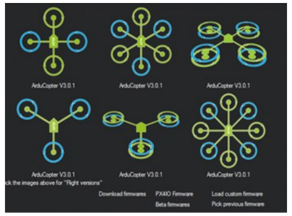

- Technical Area 1 (Explainable Learners) Milestones:
  - Demonstrate the explainable learners against problems proposed by the developers (Phase 1)
  - Demonstrate the explainable learners against common problems (Phase 2)
  - Deliver software libraries and toolkits (at the end of Phase 2)
- Technical Area 2 (Psychology of Explanation) Milestones:
  - Deliver an interim report on psychological theories (after 6 months during Phase 1)
  - Deliver a final report on psychological theories (after 12 months, during Phase 1)
  - Deliver a computational model of explanation (after 24 months, during Phase 2)
  - Deliver the computational model software (at the end of Phase 2)

# XAI Evaluation

## Challenge Problems

Analytics

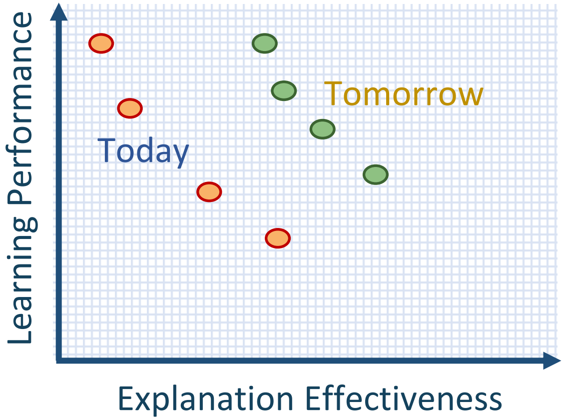

Autonomy



## Evaluation Framework

- Evaluation protocols
- Training environment
  - Training data
  - Simulation environ.
- Testing environment
  - Subjects
  - Web infrastructure
- Baseline systems

## Measurement




- **PI**: David Aha

- Justin Karneeb (Knexus)
- Matt Molineaux (Knexus)
- Leslie Smith (NRL)
- Mike Pazzani (UC Riverside)

# Phase I: Attention Map

# 1. Motivation: Being able to explain is crucial for gaining trust

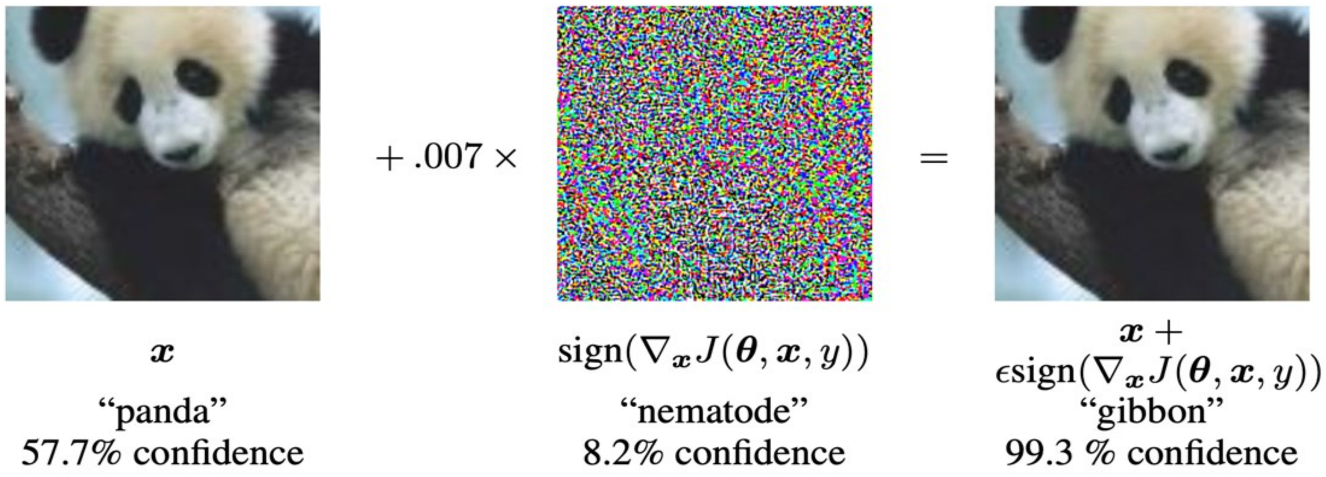
Explanation is crucial for:

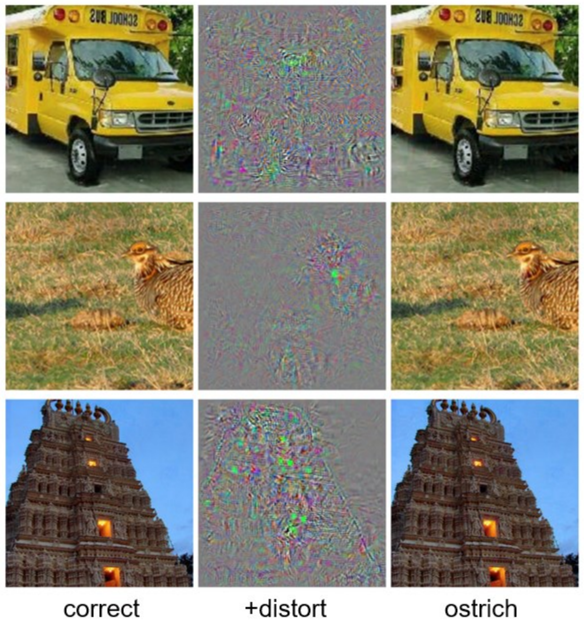
- System Safety
- Debugging
- Causality
- Justice
- Public Relations

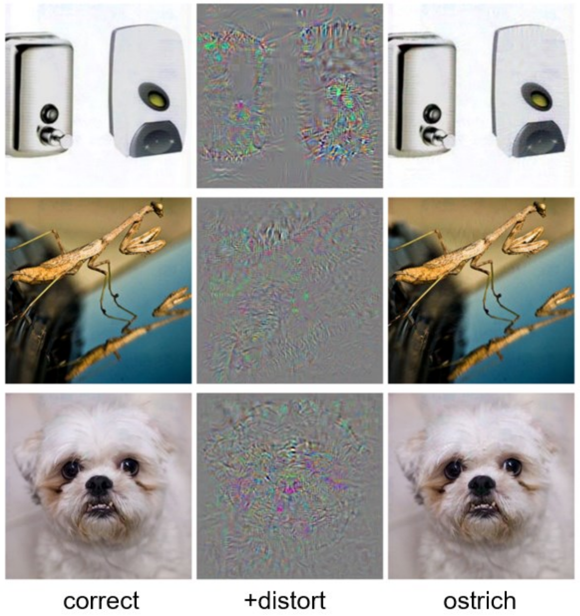
Justified Trust :
Knowing when a person system works and when does not !

**Example:** In 2015, An Amtrak Passenger Train 188 had reached a speed of 106 mph at a curve with speed limit 50 mph and derailed. 8 died and 85 were sent to hospital. The public perspective on the train driver drastically changed from The driver being “absolutely guilty” to “not really his fault” after explaining them the causes.

# A crisis for deep models: crucial applications cannot trust DNN models





[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." *ICLR 2014.*
[2] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *ICLR 2015.*

# 2. Concepts:  What are Interpretation and Explanation?

**Interpretable representation:**

Establish a common language between machines and humans.

A language is the set produced by a grammar (And-Or Graph).

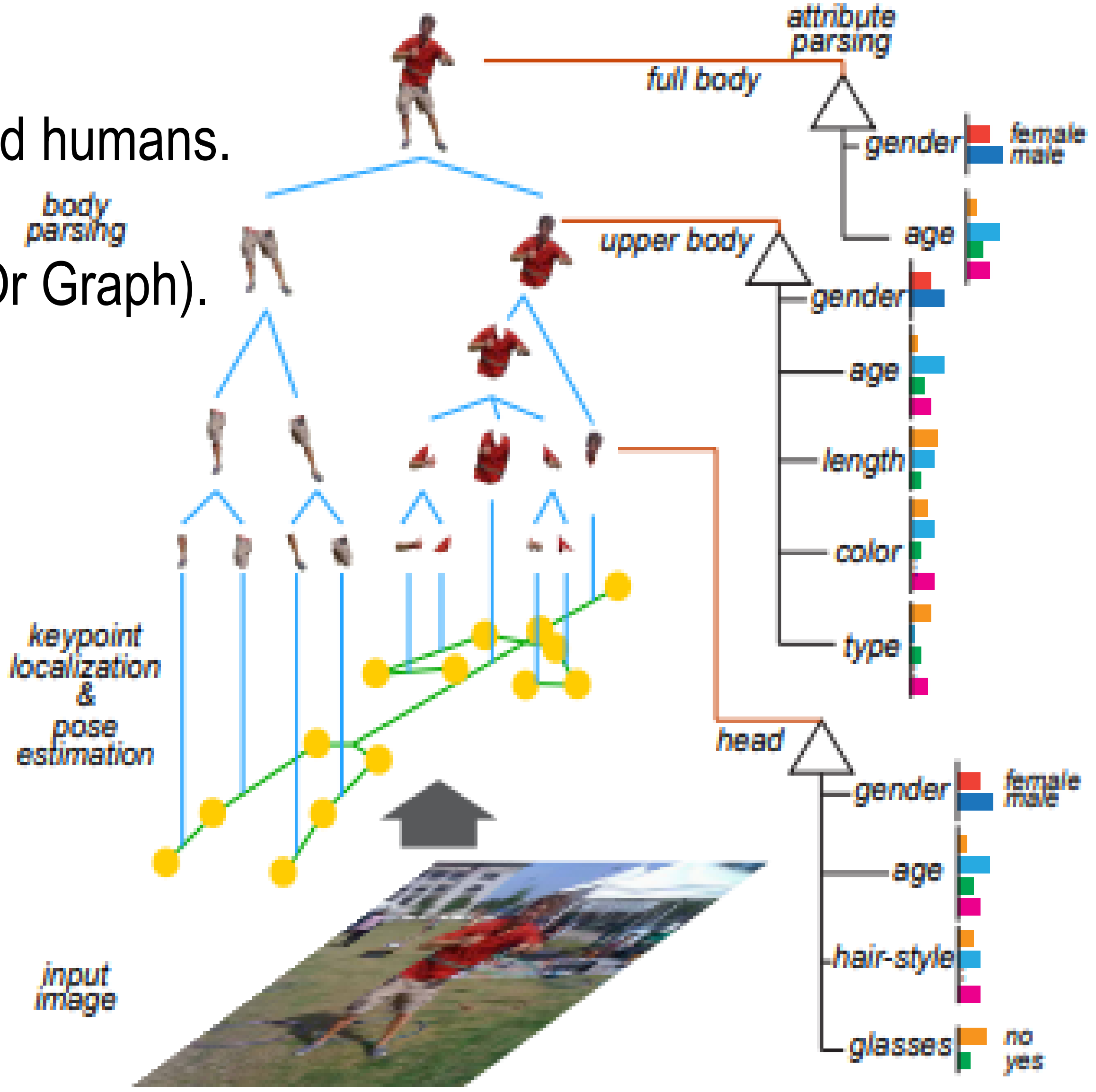
**Example: attribute parse graph**

**Appearance attributes**:

- Male/female;
- Clothes style;
- Hair styles etc;
- Hat/Glass/…

**Geometric Attributes**:

- Pose and parts
- Actions
- Interactions with objects.

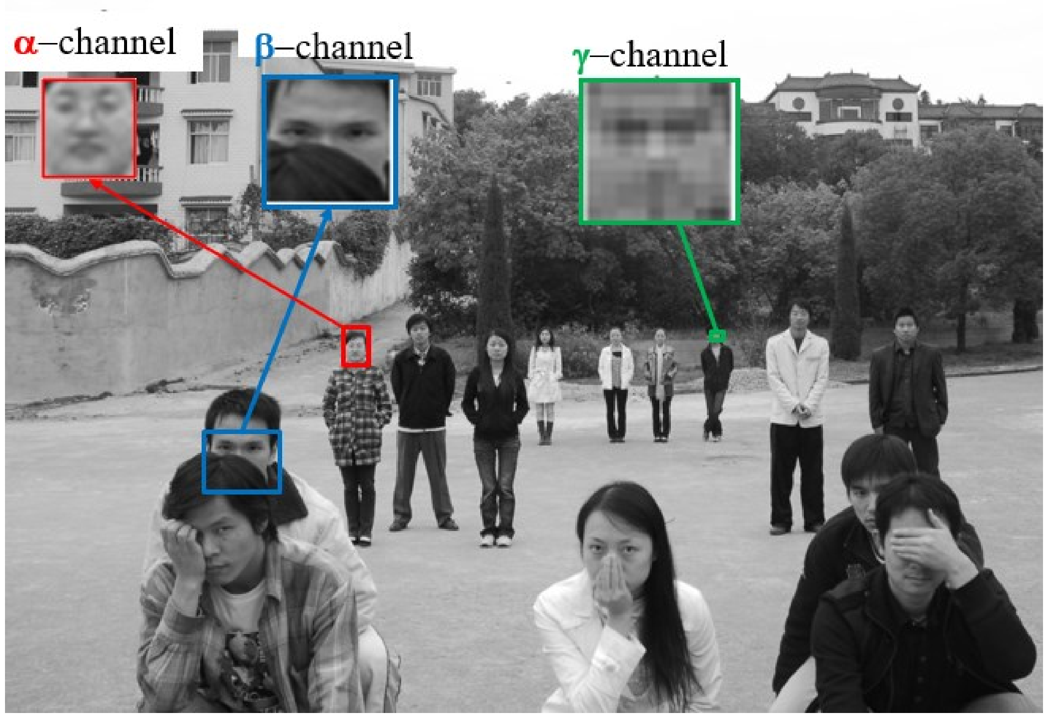


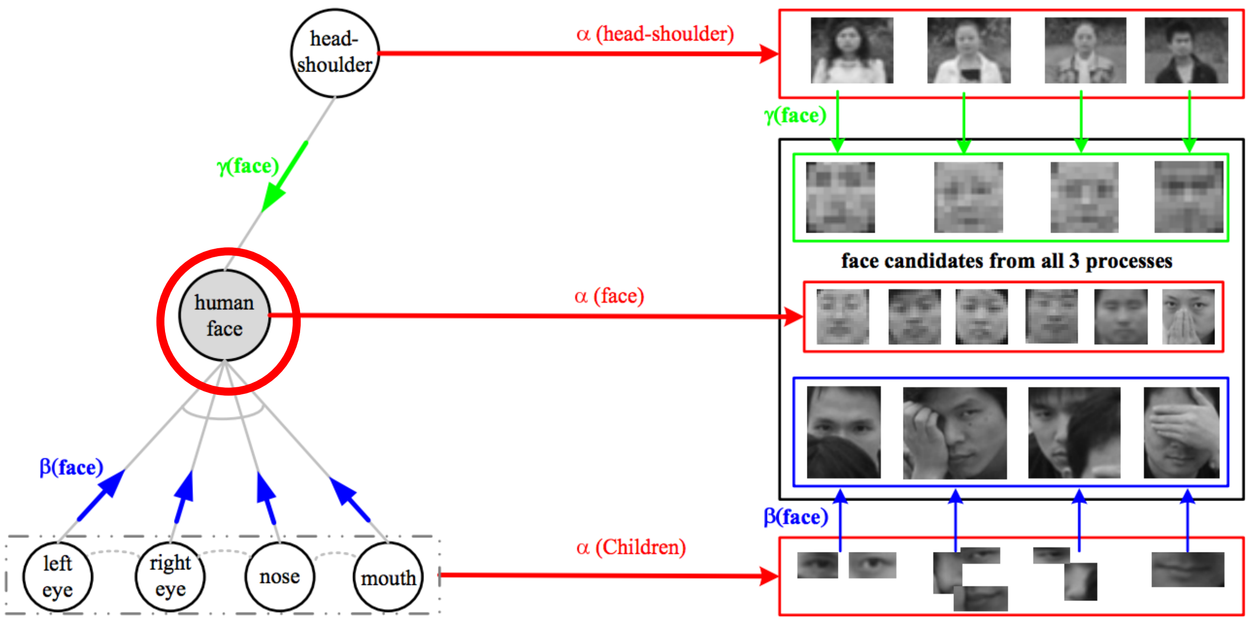

Seyoung Park, Xiaohan Nie, Brandon Rothrock, Song-Chun Zhu

# Explanation is built on an interpretable representation

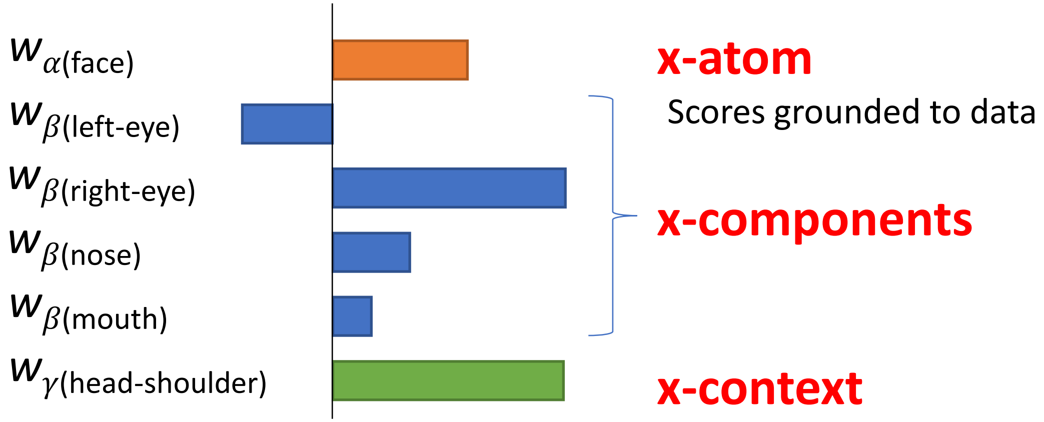
Example: Why is it a Human Face?

$\alpha$-$\beta$-$\gamma$ pathways for recognition.



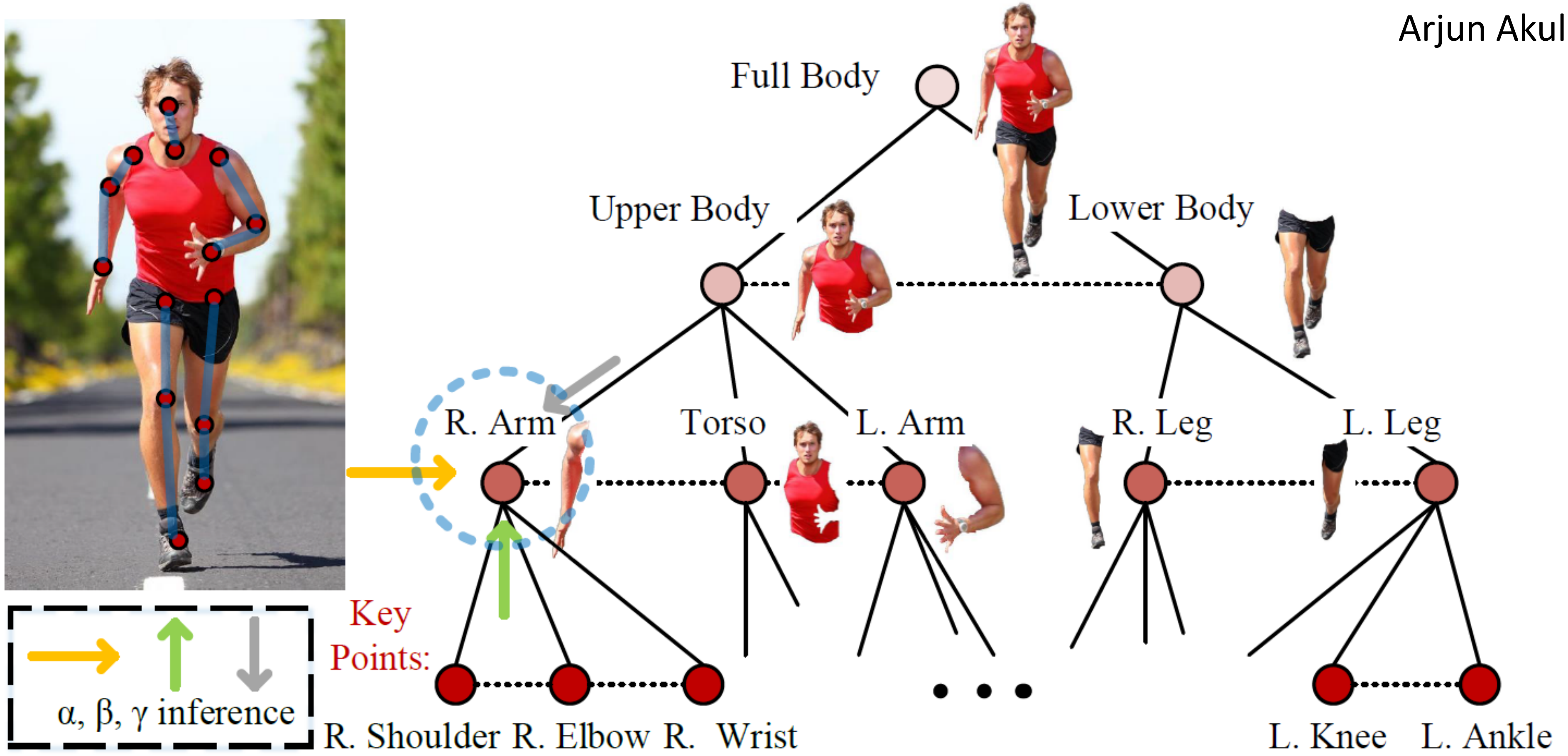

Explanations from all paths




Tianfu Wu, Song-Chun Zhu

# Recursive $\alpha$-$\beta$-$\gamma$ channels in a parse graph

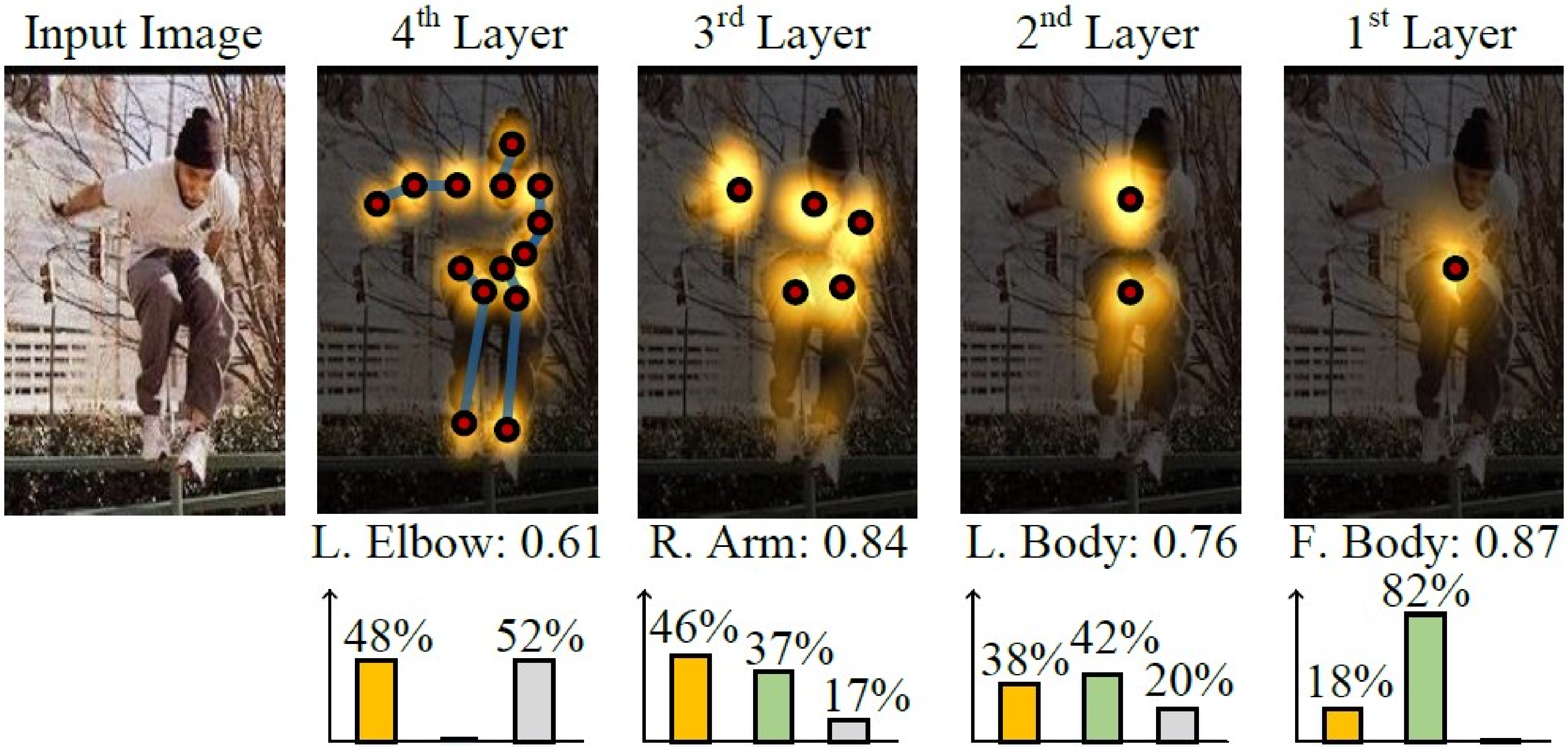
Arjun Akula, Song-Chun Zhu




## Interpretability = entropy ( prob.(parse graph | input image) ).

For most daily images, we usually perceive only 1 interpretation. Otherwise we are confused all the time.
This is because we stop growing the parse graph when the entropy is too high, as it becomes speculation

# Example: Calculating the contributions of $\alpha$-$\beta$-$\gamma$ channels



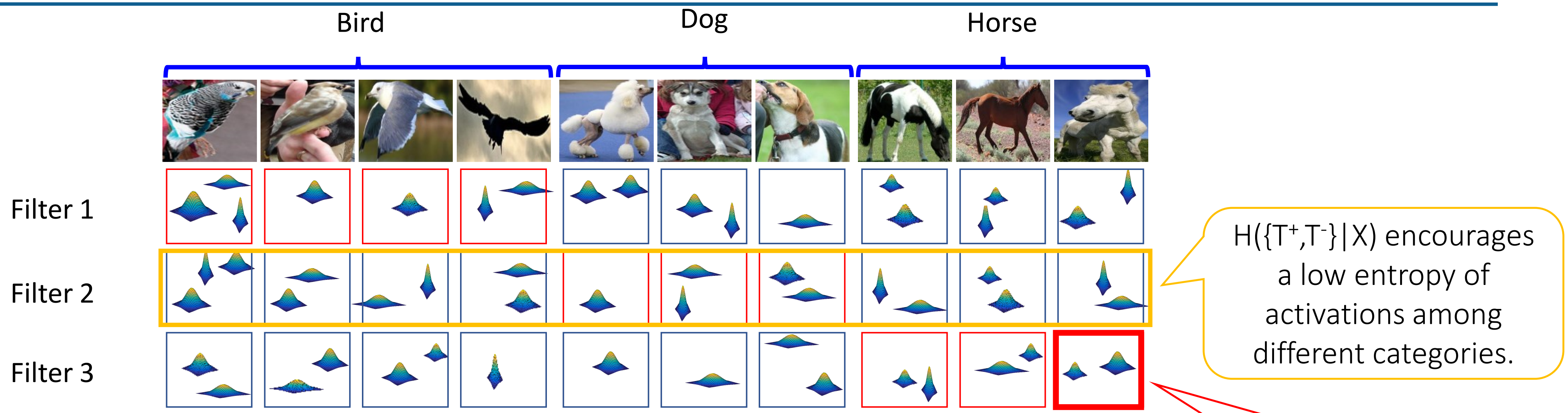

Prediction maps from different layers

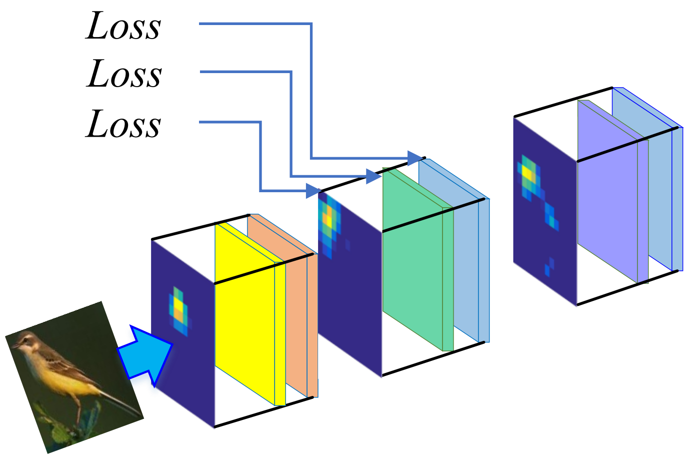
Prediction score

Contribution of each channel.

Wenguang Wang, Song-Chun Zhu

# DNN is not interpretable, as its neurons have “many-to-many” mapping to categories.

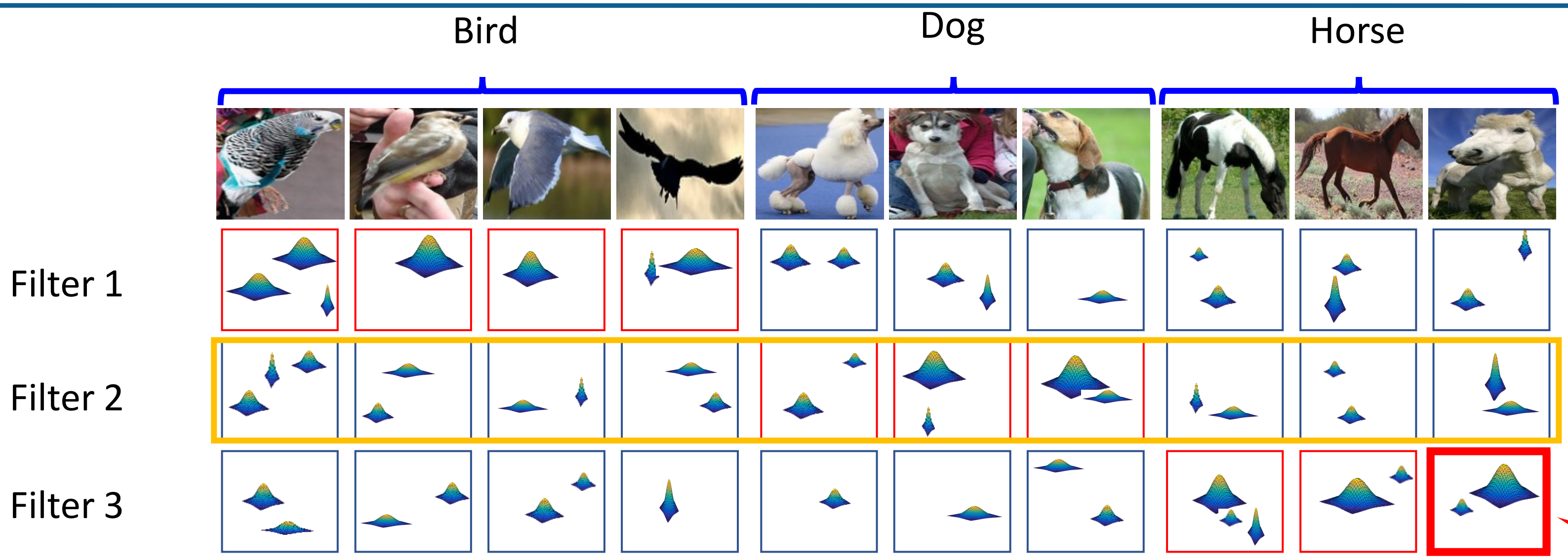



$$\mathbf{Loss}_f = -H(\mathbf{T}) + H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X}) + \sum_x p(\mathbf{T}^+, x) H(\mathbf{T}^+|X = x)$$

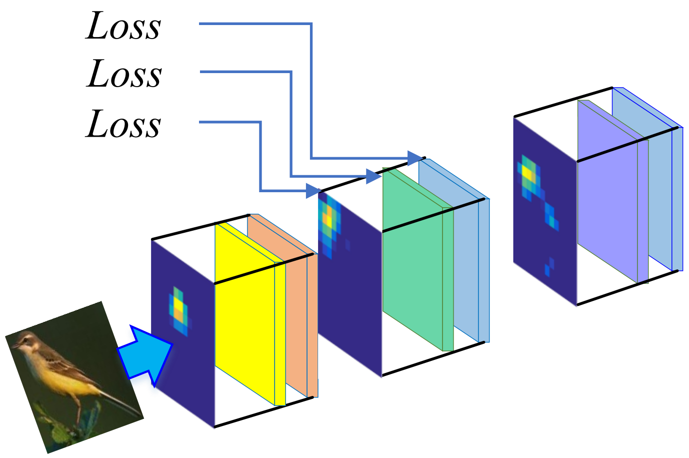
Quanshi Zhang, Ying Nian Wu, Song-Chun Zhu

# Adding regularization term to minimize the entropy of interpretation




H({T⁺,T⁻}|X) encourages a low entropy of activations among different categories.

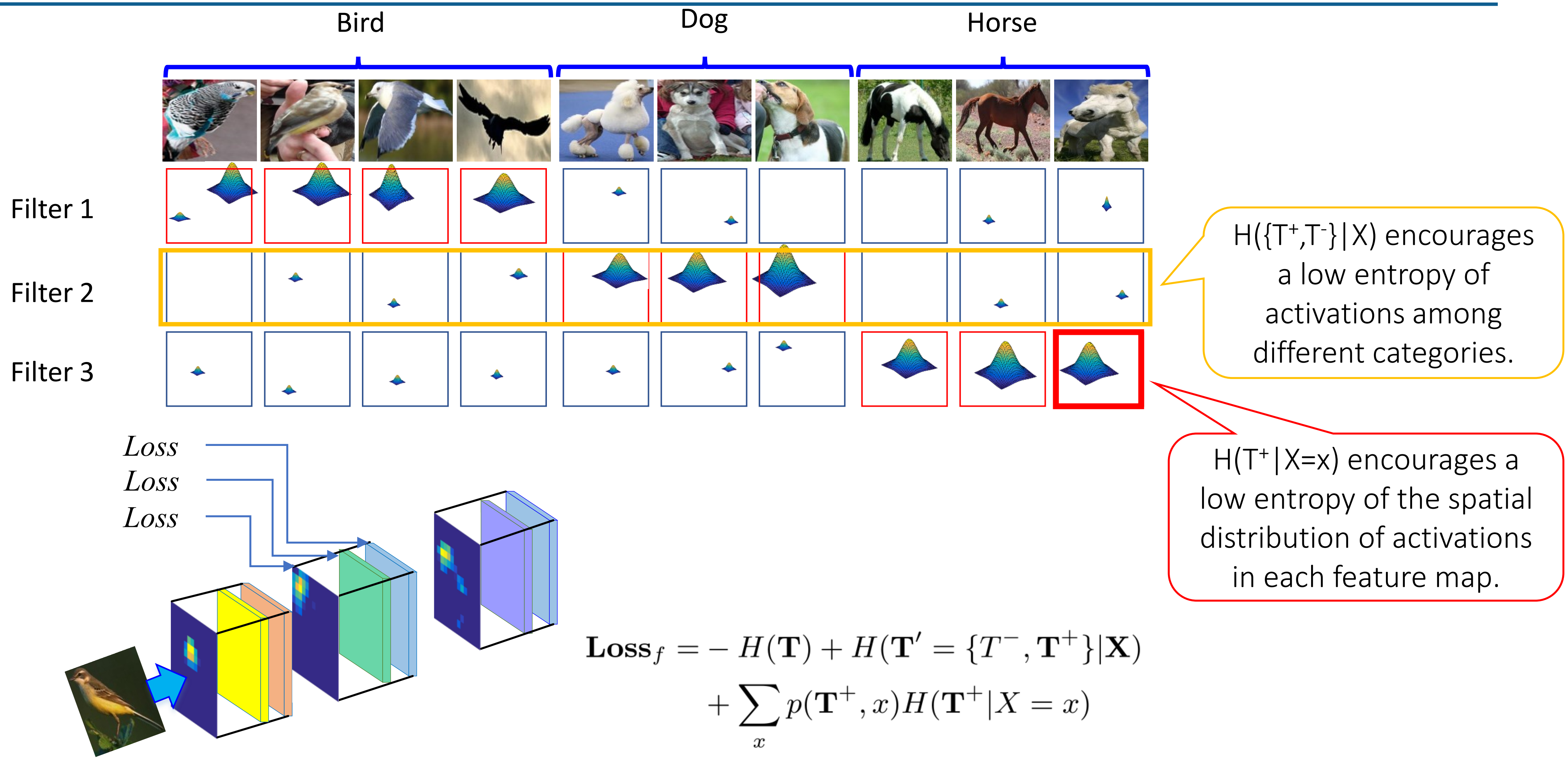
H(T⁺|X=x) encourages a low entropy of the spatial distribution of activations in each feature map.




$$\mathbf{Loss}_f = -H(\mathbf{T}) + H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X}) + \sum_x p(\mathbf{T}^+, x) H(\mathbf{T}^+|X = x)$$

Quanshi Zhang, Ying Nian Wu, Song-Chun Zhu

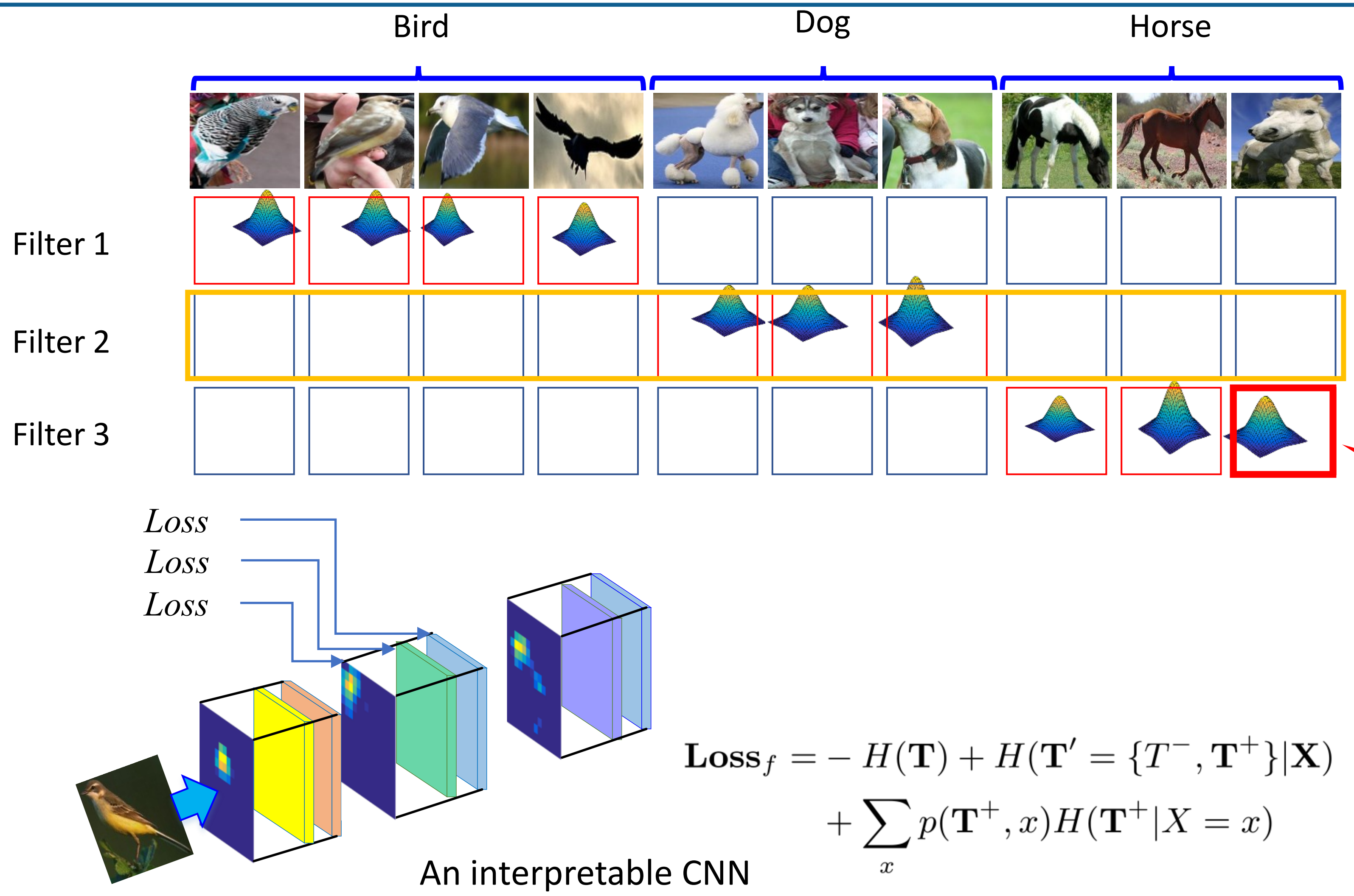
# Adding regularization term to minimize the entropy of interpretation




$$\mathbf{Loss}_f = -H(\mathbf{T}) + H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X}) + \sum_x p(\mathbf{T}^+, x) H(\mathbf{T}^+|X = x)$$

Quanshi Zhang, Ying Nian Wu, Song-Chun Zhu

# Adding regularization term to minimize the entropy of interpretation !




H({T⁺,T⁻}|X) encourages a low entropy of activations among different categories.

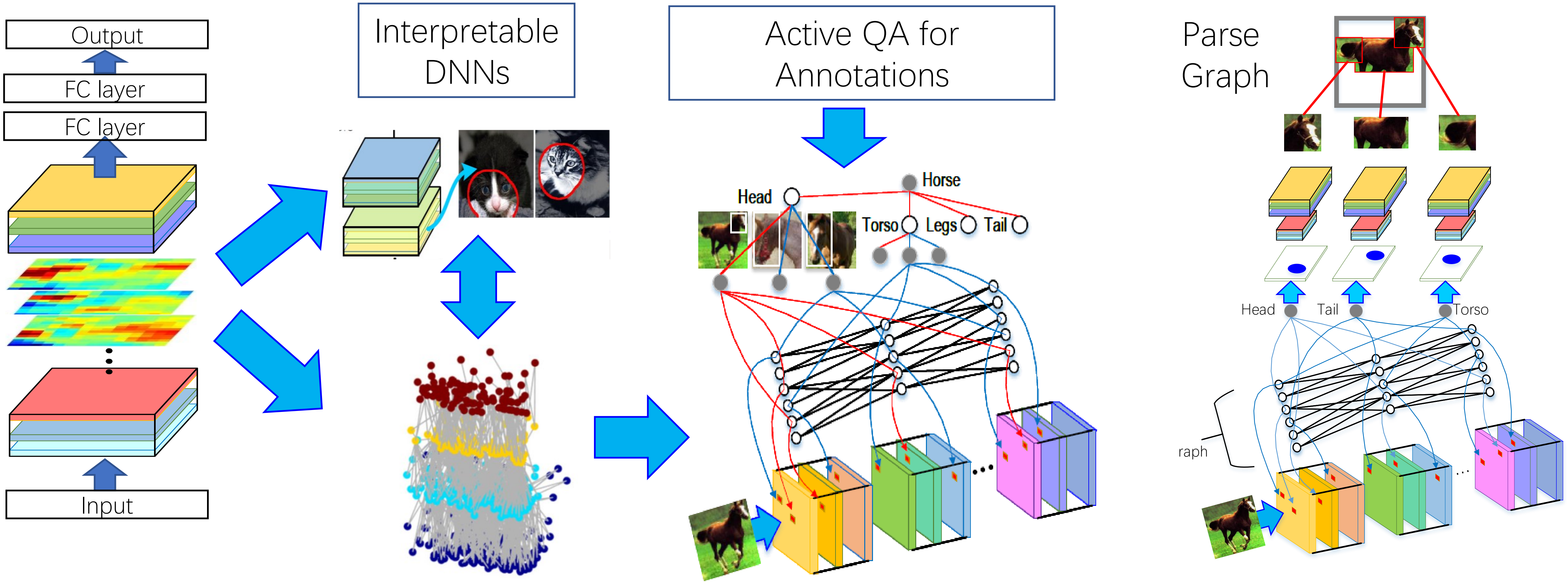
H(T⁺|X=x) encourages a low entropy of the spatial distribution of activations in each feature map.

$$\mathbf{Loss}_f = -H(\mathbf{T}) + H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X}) + \sum_x p(\mathbf{T}^+, x) H(\mathbf{T}^+|X = x)$$

Quanshi Zhang, Ying Nian Wu, Song-Chun Zhu

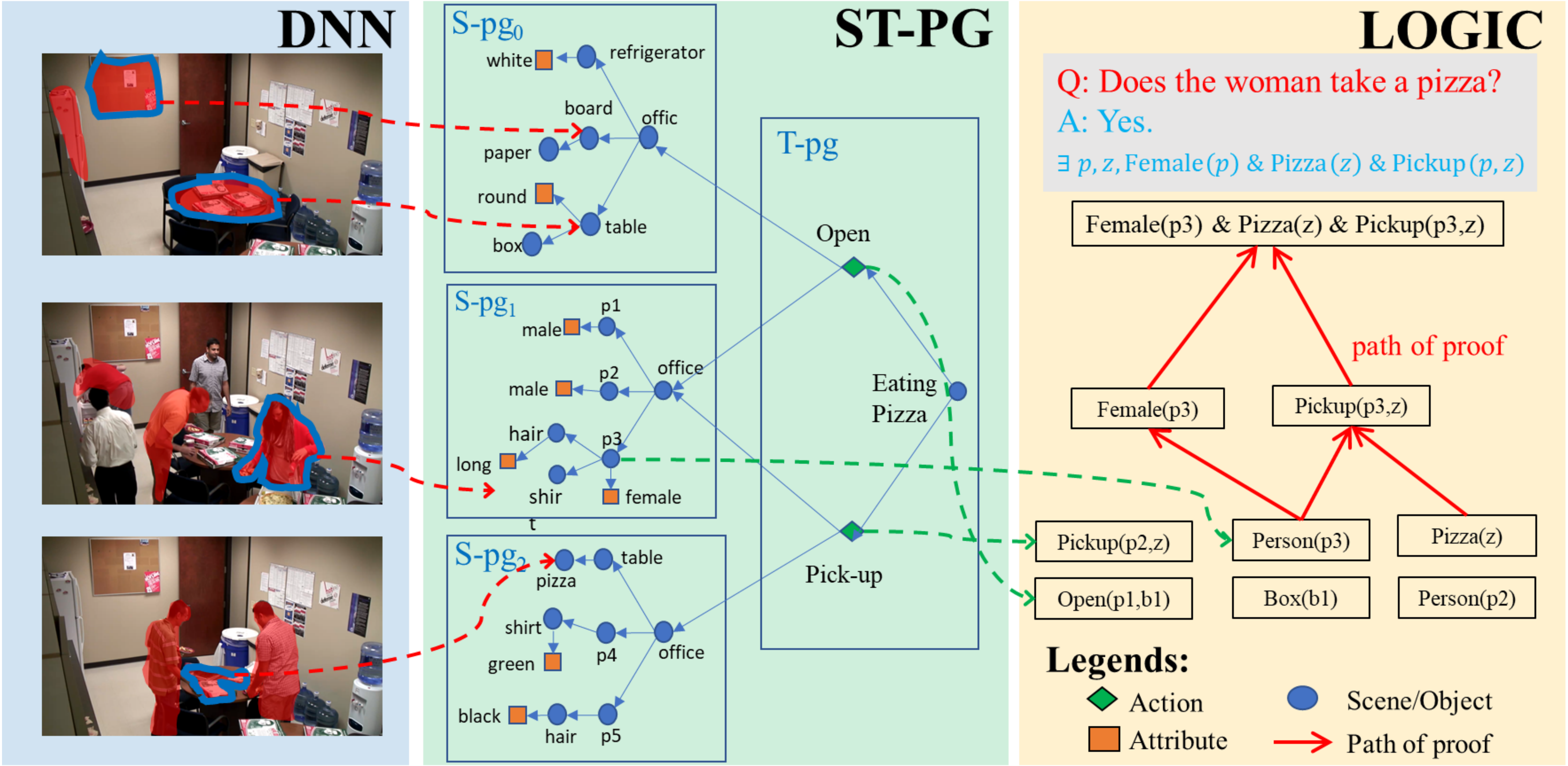
# Disentangle DNN neurons into an Interpretable DNNs

Disentangle DNN neurons, and map them to nodes in parse graph.

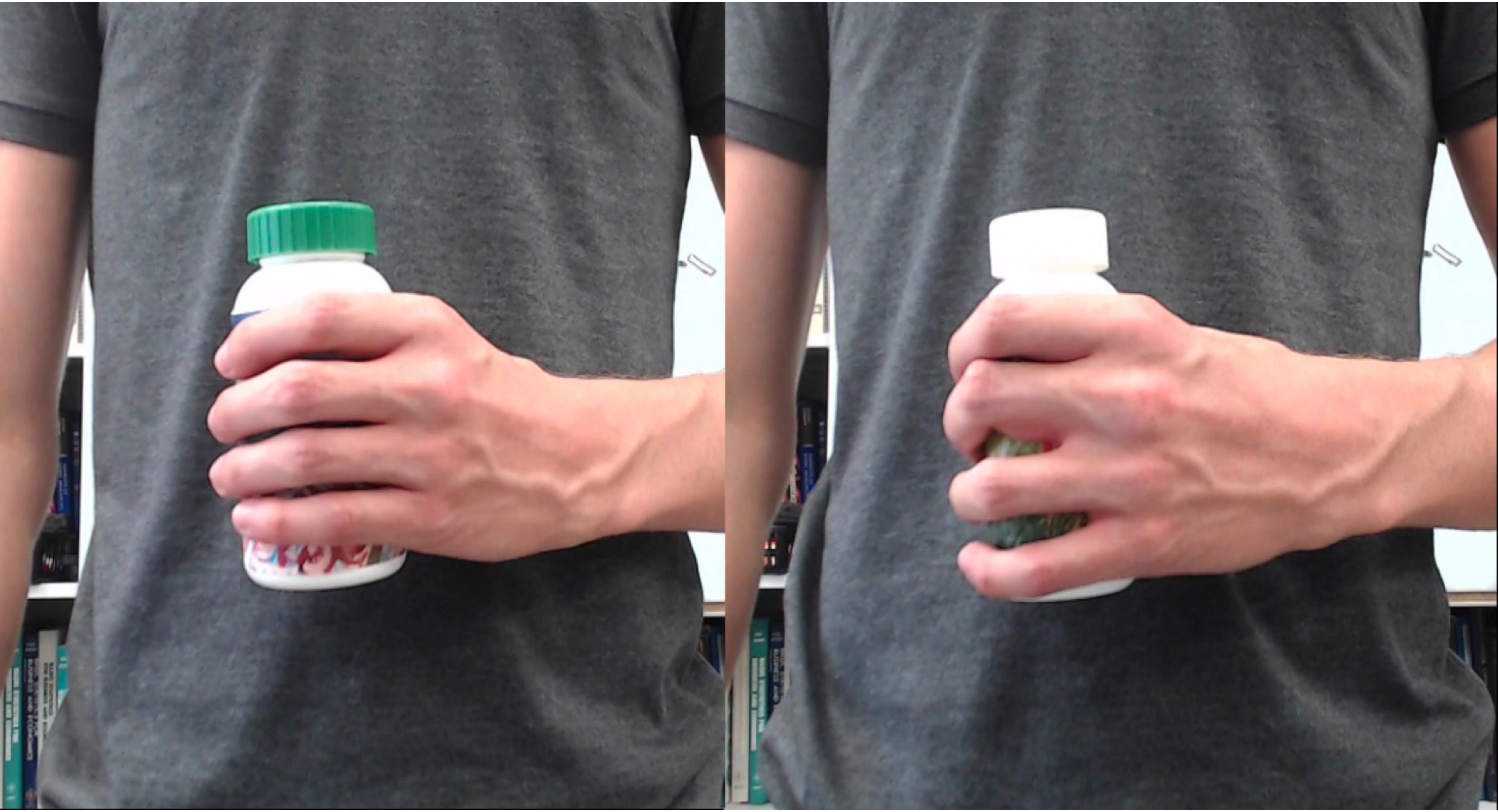


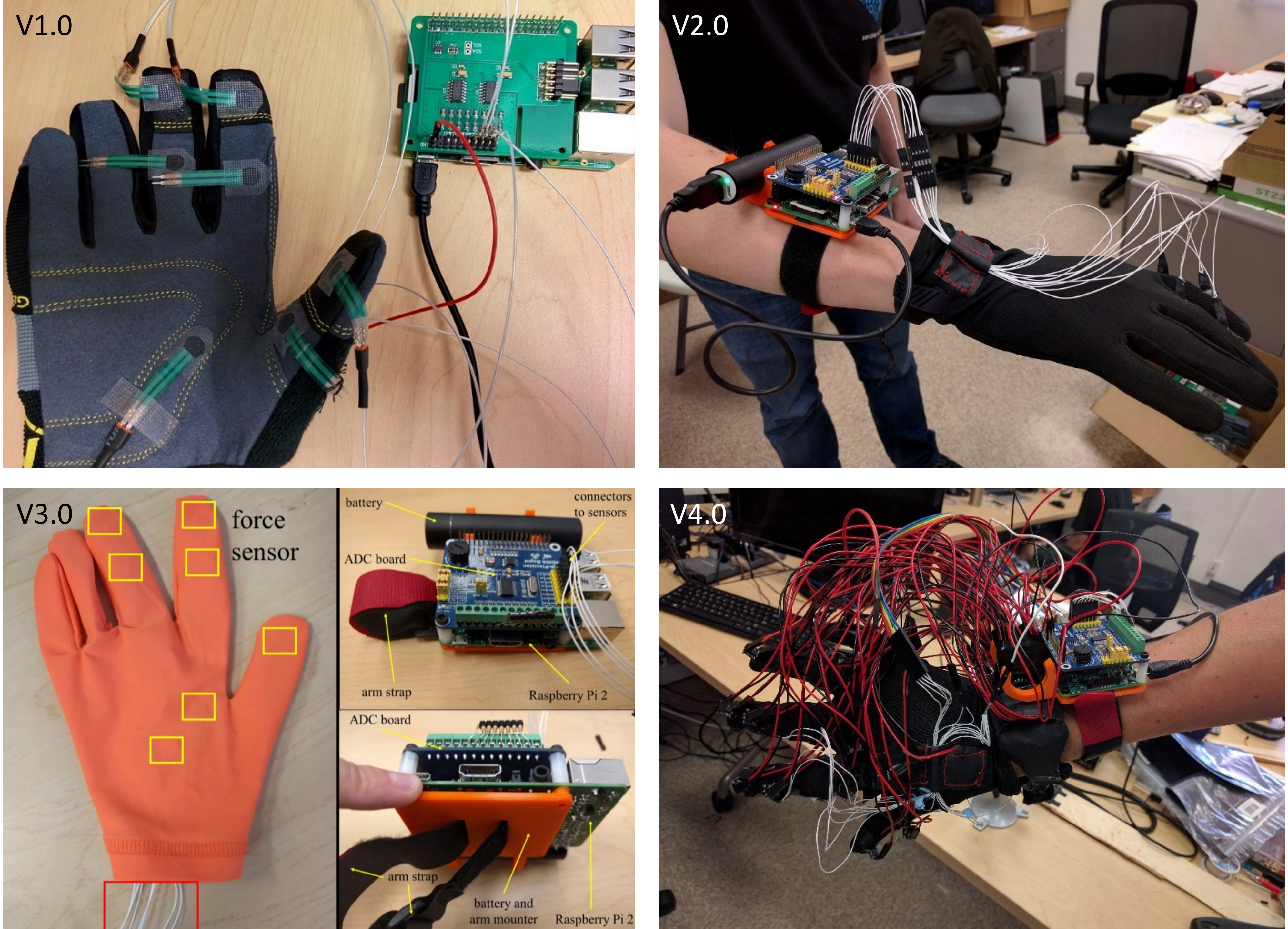

Quanshi Zhang, Ying Nian Wu, Song-Chun Zhu

# Close the Loop: DNN – AOG – LOGIC

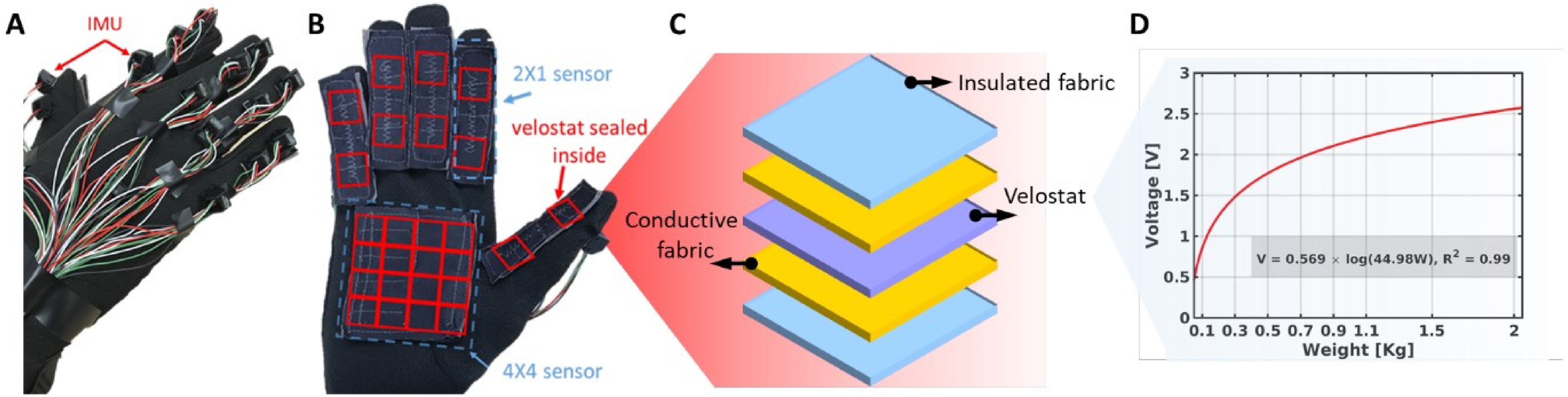


# Phase II: Enhancing Trust



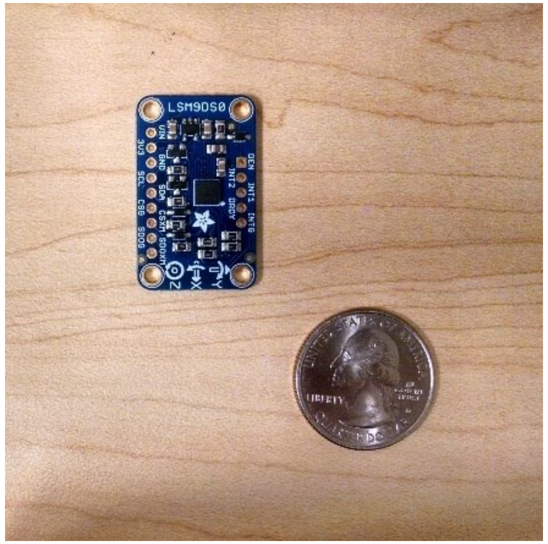

V1.0
V2.0
V3.0
force
sensor
battery
connectors
to sensors
ADC board
arm strap
Raspberry Pi 2
ADC board
arm strap
battery and
arm mounter
Raspberry Pi 2
V4.0

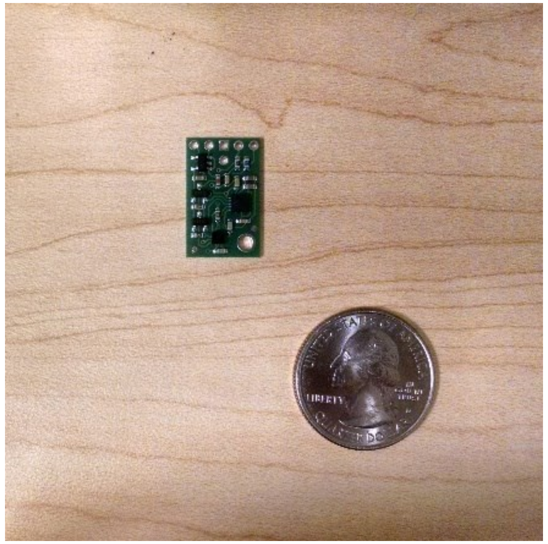
# Design of a Glove-based System

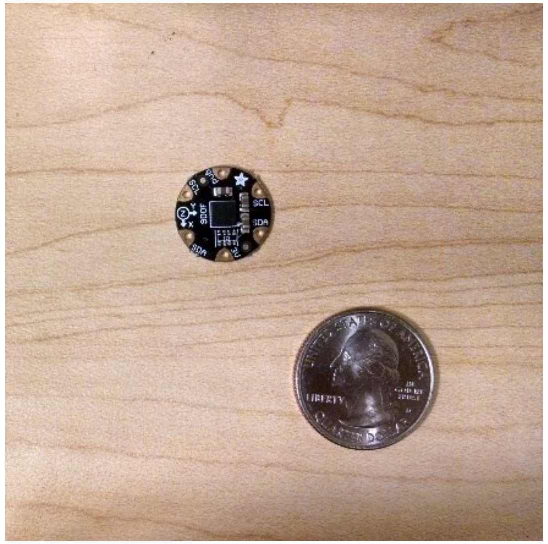


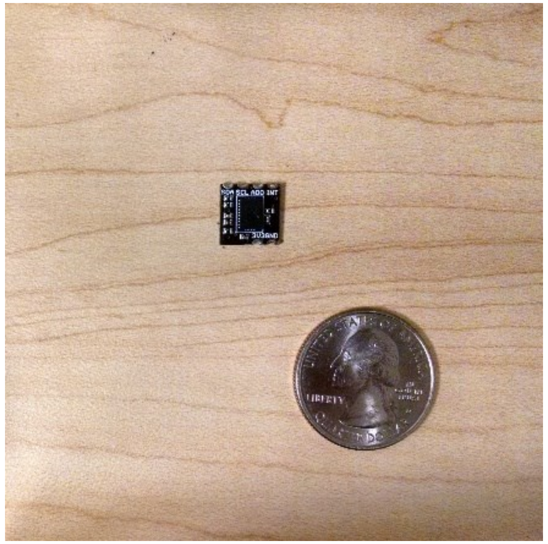

Hangxin Liu, etal

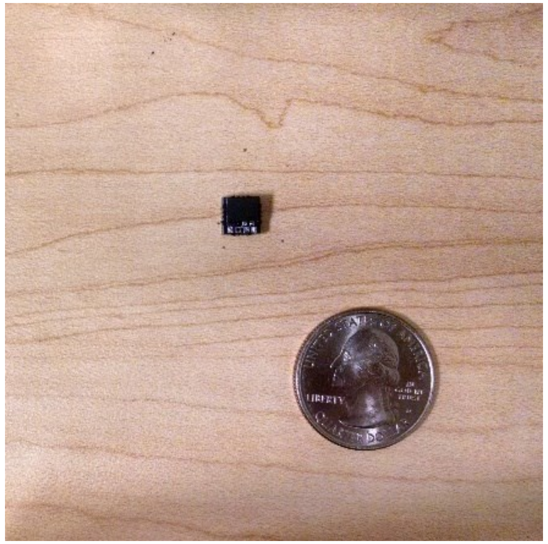
# Iteration of IMUs



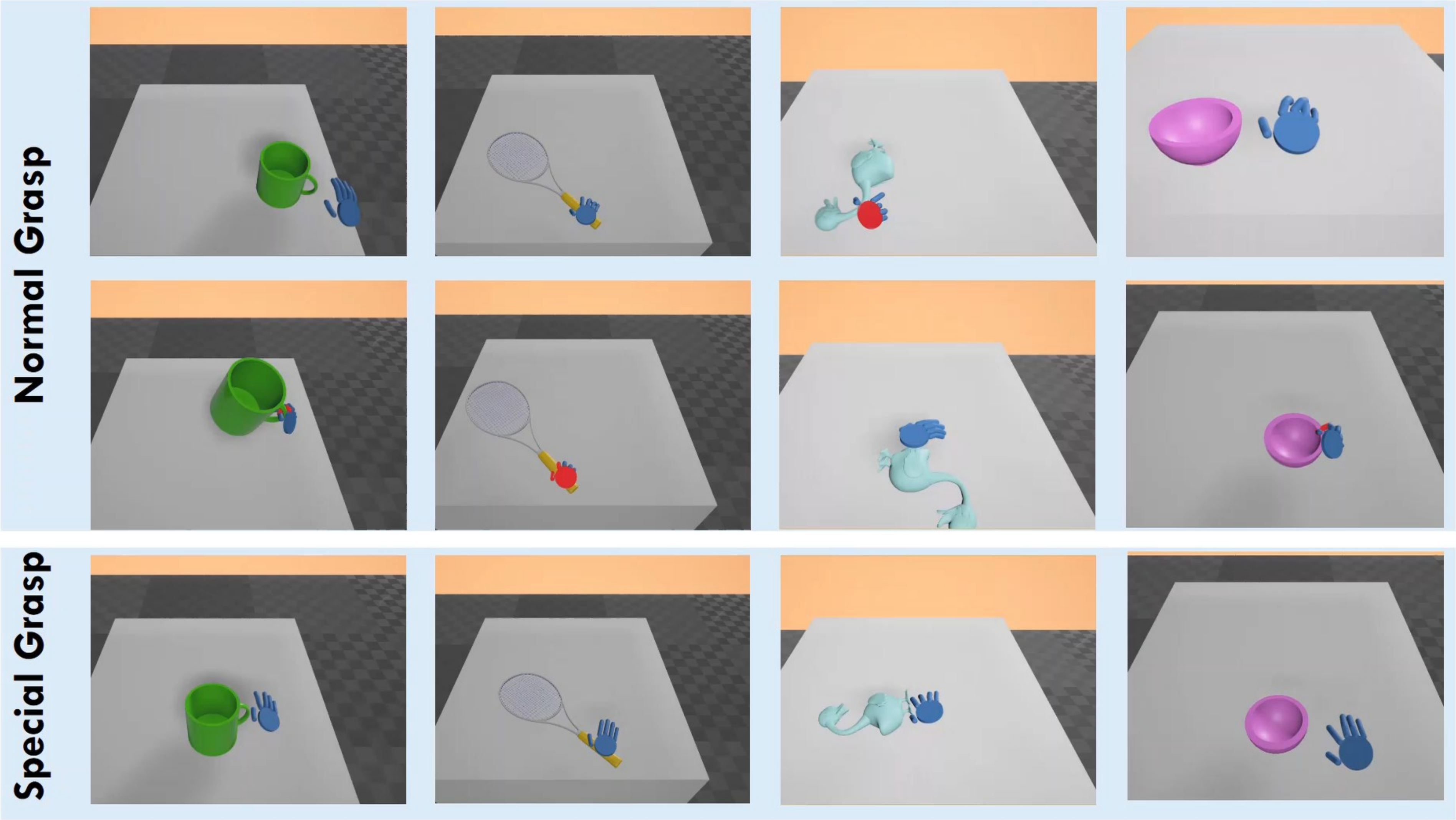




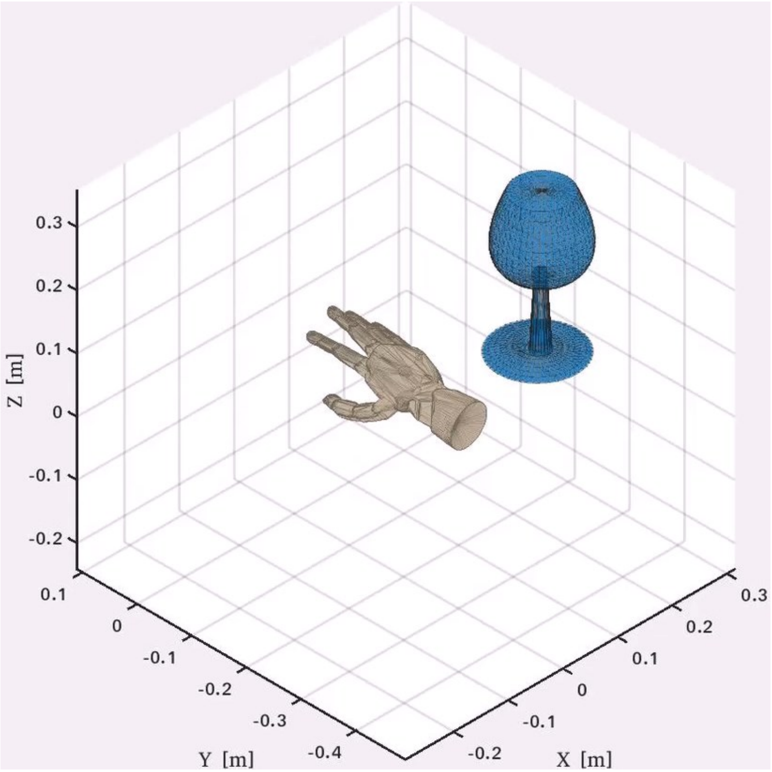


Hangxin Liu, etal

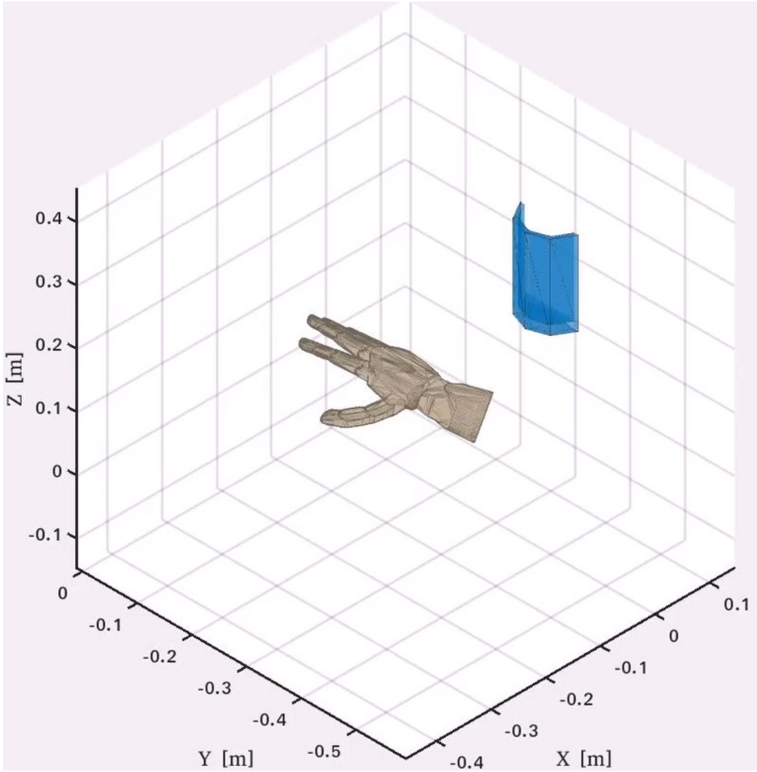
# High-Fidelity Grasping in Virtual Reality




Zhengliang Zhang, etal



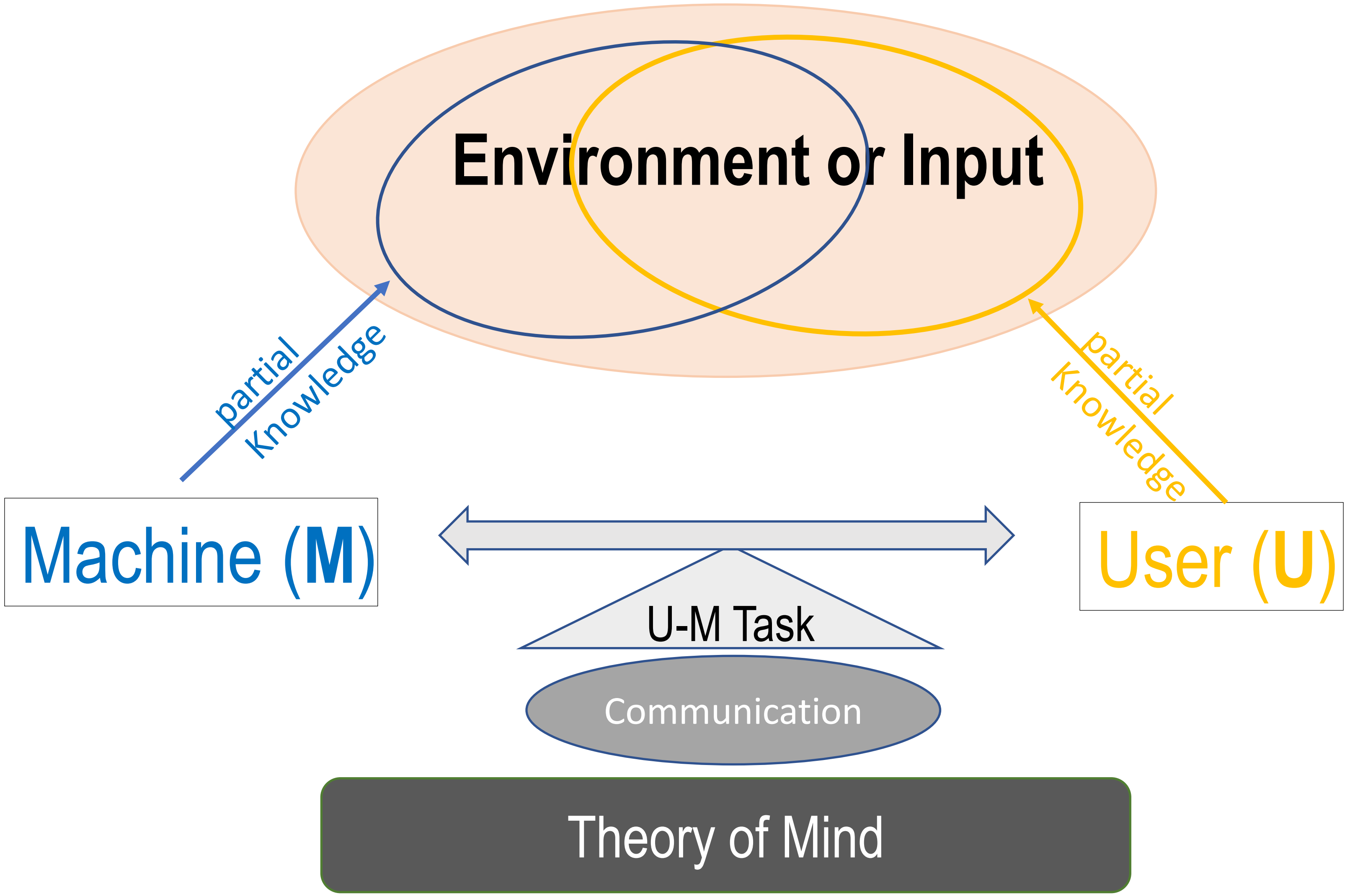
Z [m]
0.3
0.2
0.1
0
-0.1
-0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
Y [m]
-0.2
-0.1
0
0.1
0.2
0.3
X [m]

Z [m]
0.4
0.3
0.2
0.1
0
-0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
Y [m]
-0.4
-0.3
-0.2
-0.1
0
0.1
X [m]

# Phase III: Bidirectional Value Alignment

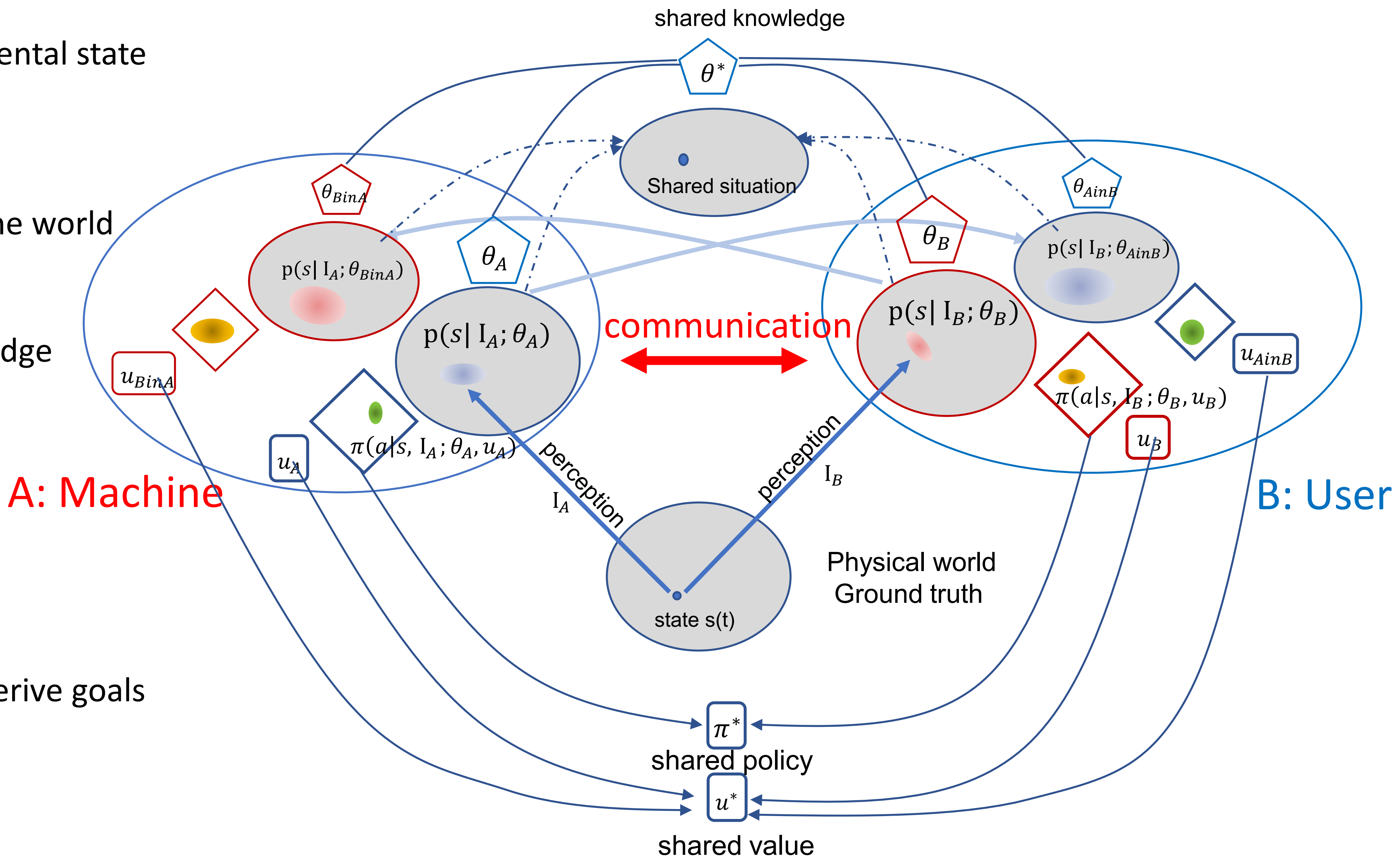
# Bidirectional Alignment in Human-Robot Collaboration



# Cognitive Architecture for Human-Machine Communication

Each ellipse represents a mental state and has 4 components:

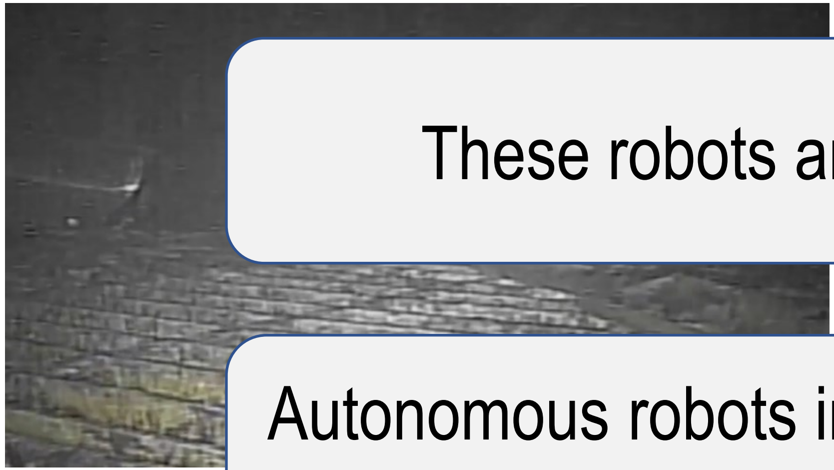
1. **Belief**
   – perceived states of the world
2. **Model**
   – concepts and knowledge
3. **Policy**
   – action and plans
4. **Value**
   – gains and losses, to derive goals

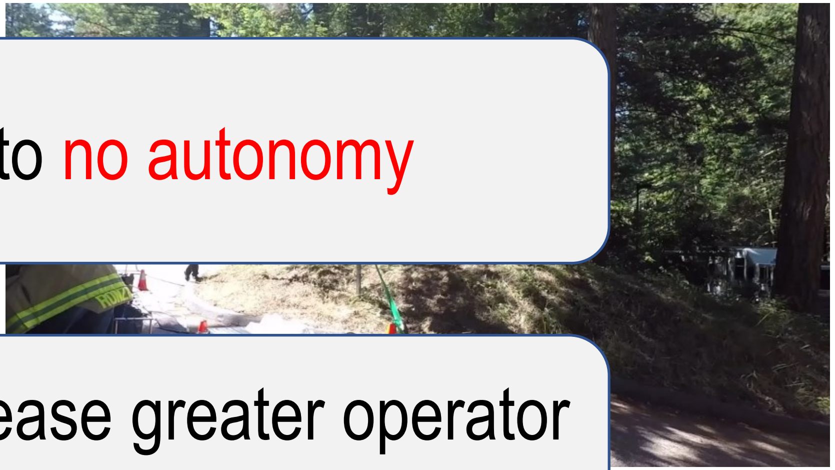


# Scenarios Requiring Bidirectional Human-Robot Value Alignment

Entering radioactive area

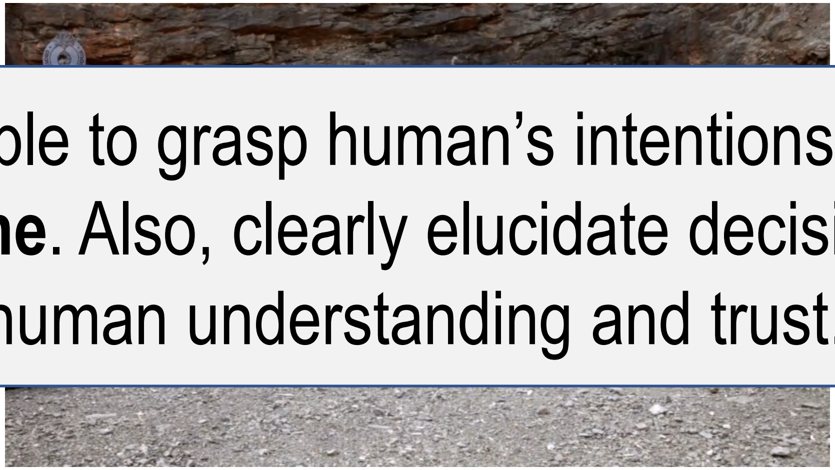
Detecting hazards





These robots are teleoperated with little to no autonomy

Autonomous robots in these domains will increase greater operator flexibility and mission scale

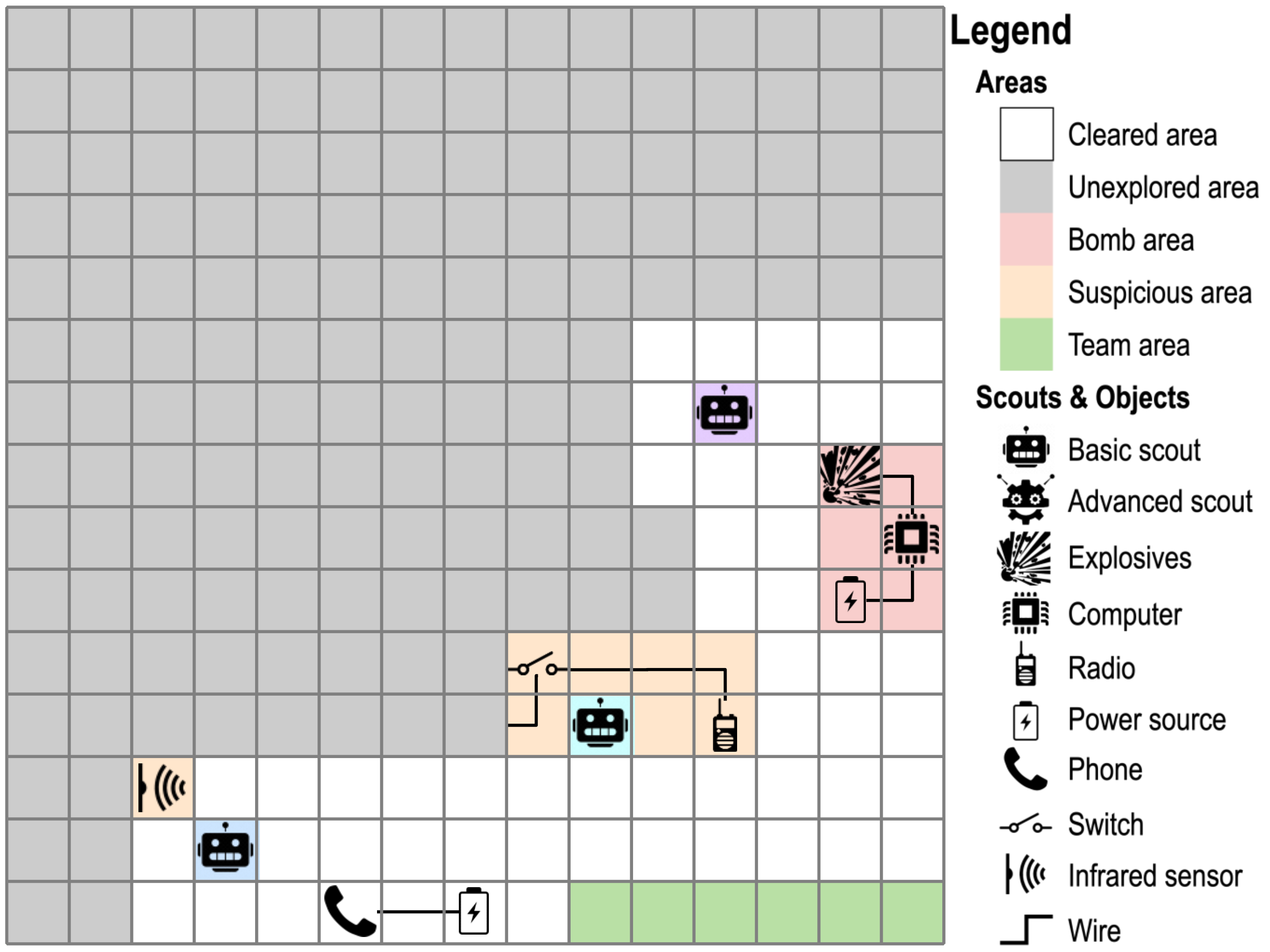


Robots must be able to grasp human’s intentions and values of the task **in real-time**. Also, clearly elucidate decision-process for human understanding and trust.

# Prototypical Setting: Scout Exploration Game

## User-machine task setting

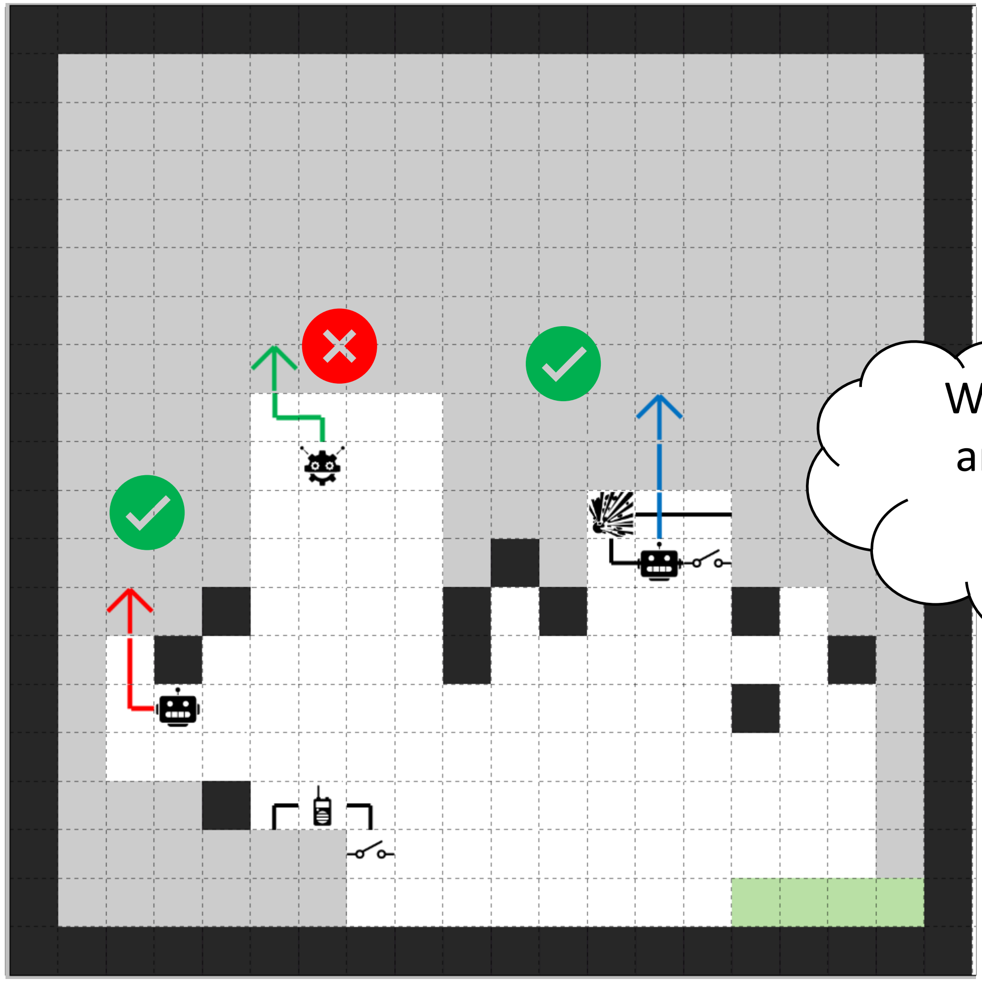
- A human-robot team is trying to find a safe path crossing an unknown terrain from the bottom right to the top left
- Additional goals may be achieved:
  - Find the path as fast as possible
  - Collect extra resources
  - Defuse bombs in the map
  - Detect as much region as possible

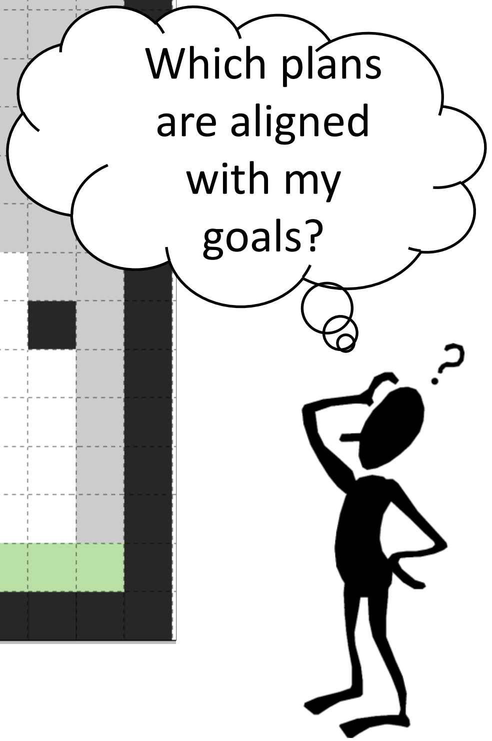


# Prototypical Setting: Scout Exploration Game

## User-machine task setting

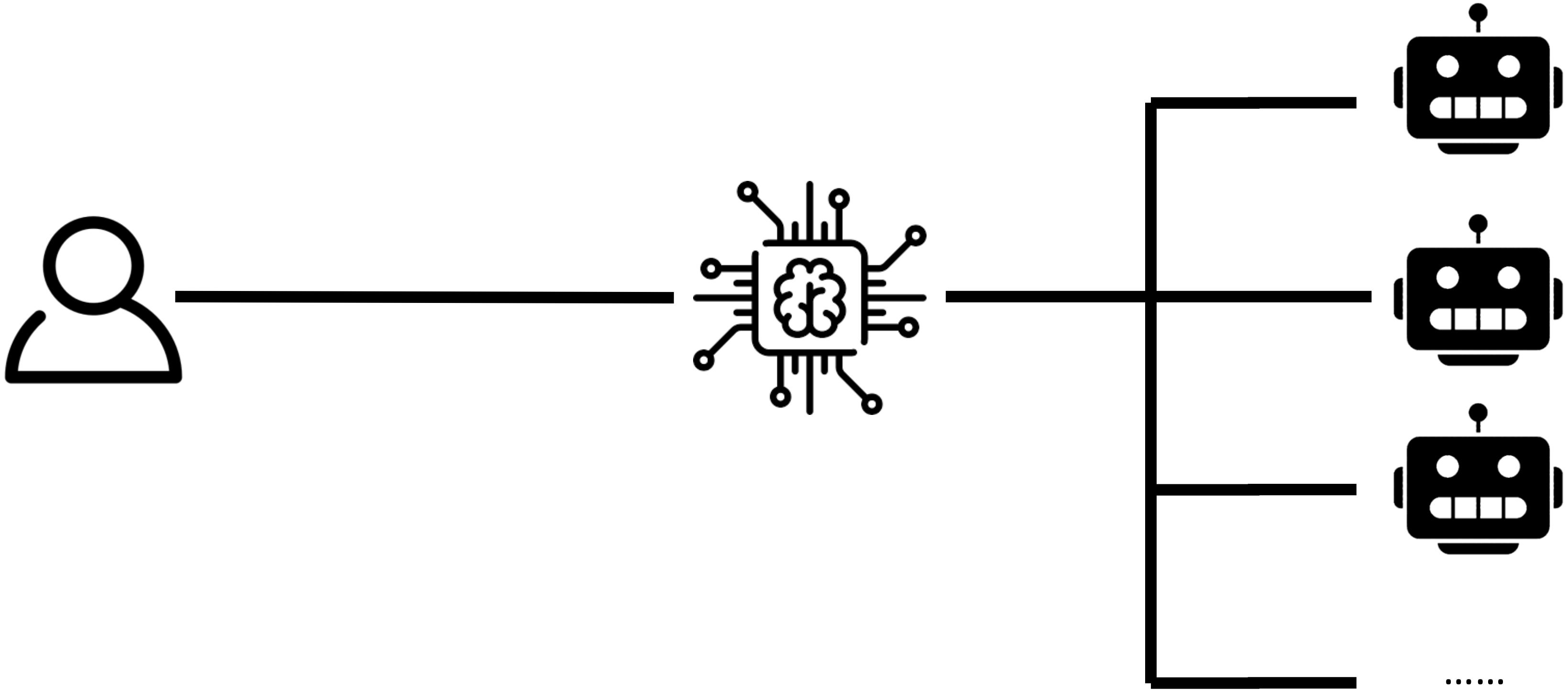
- The robots act as scouts to explore potential bombs and communicate with user.
- User can accept or reject proposals from the robot scouts
- Requires bidirectional human-robot alignment
  - Understanding human values by proposals
  - Elucidate self by providing proper explanations





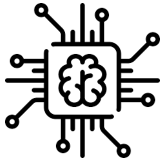
# Task Specification

- Only the scouts are interacting with the physical states via actions/observations
- Human has hidden information that the scouts need to finish the task
- Human can only interact with a centralized agent which controls all the scouts

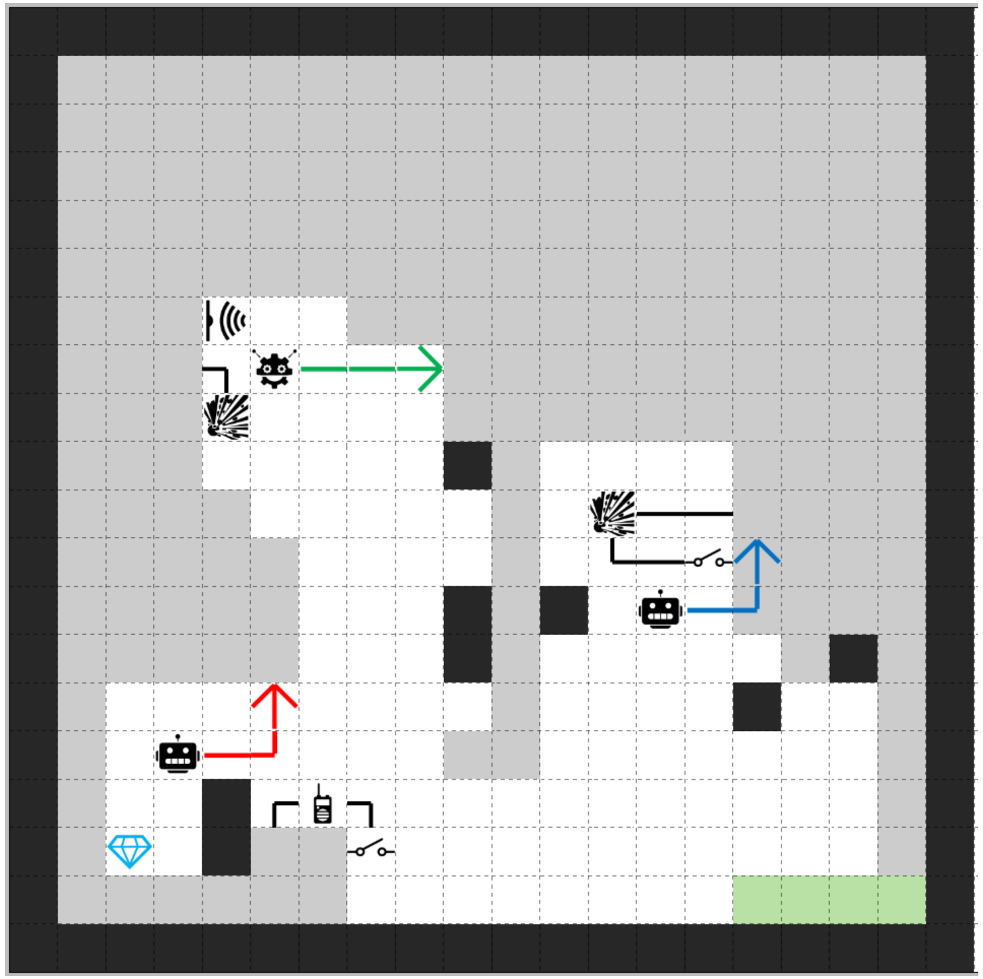


# Scout Exploration Game Design

- Infer the importance of the goals and values through communication with the human



- Control scouts to interact with the environment via action & observation

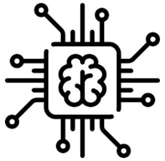


- Knows the importance of the goals

- Knows about the map via robots' messages & instructs robots to act

# Scout Exploration Game Design

- Can only infer the importance of the goals through communication with the human



- Directly control scouts to interact with the environment via action & observation

Inform about the state

Inform about goals and value

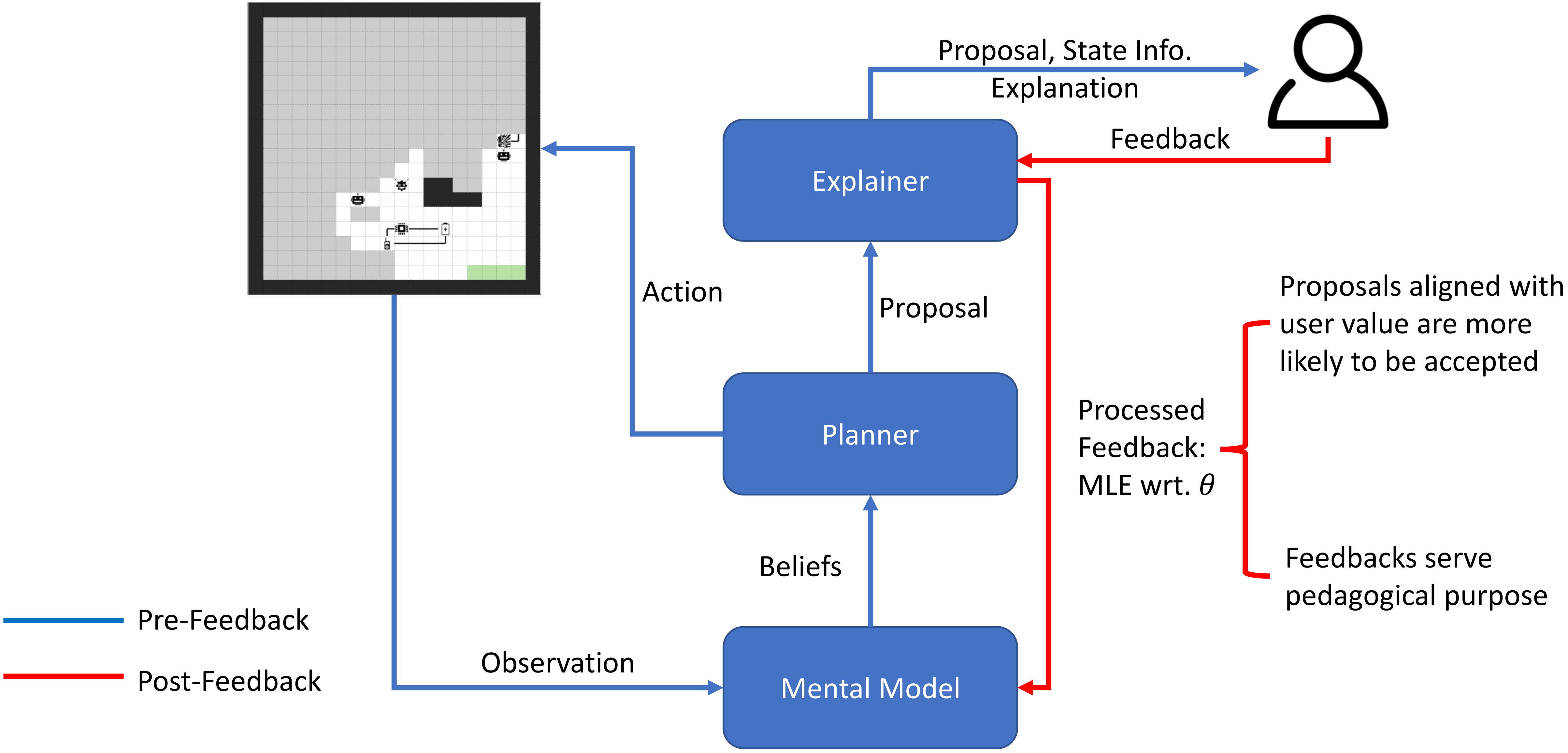
- Knows the importance of the goals

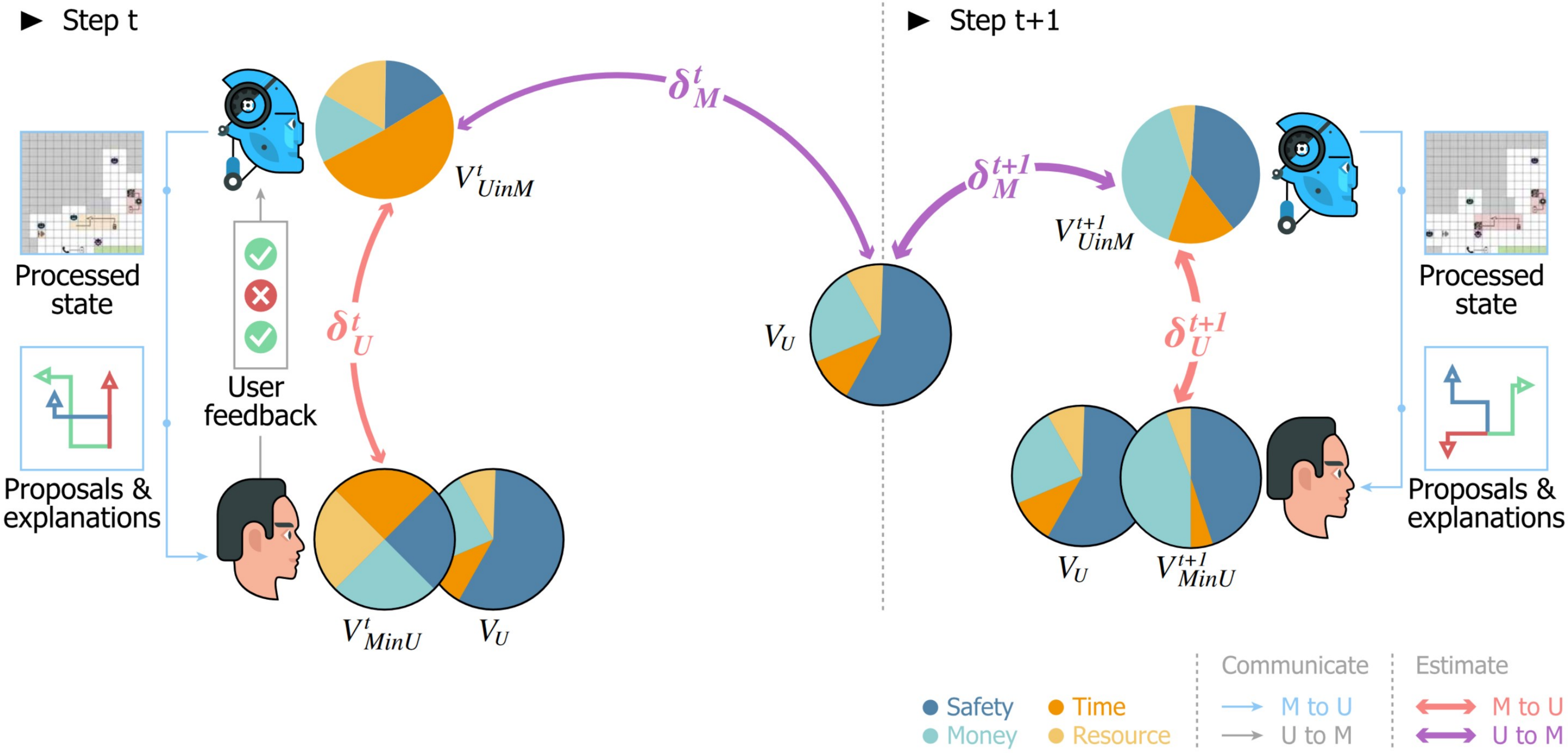
- Knows about the map via robots' messages & instructs robots to act

# The Need for Explanation

- Asymmetric information between human and robot
  - Robots have access to additional sensing information
  - Human has access to value function
- Scouts providing state information → high human cognitive burden
- Scouts providing actions proposals → some cognitive relief
- Scouts providing explanations → greater cognitive relief
- Improving user-machine task performance, and scaling up the team.

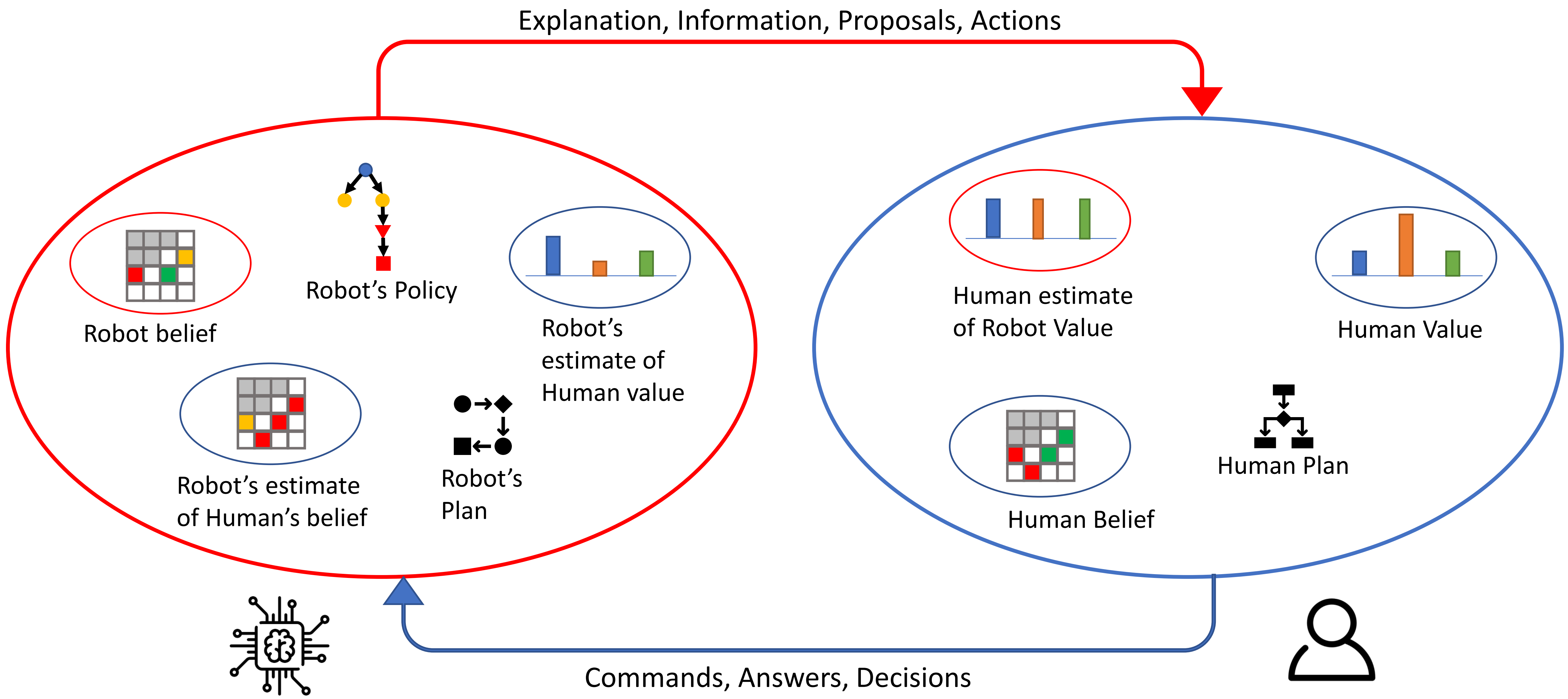
# Computational Framework

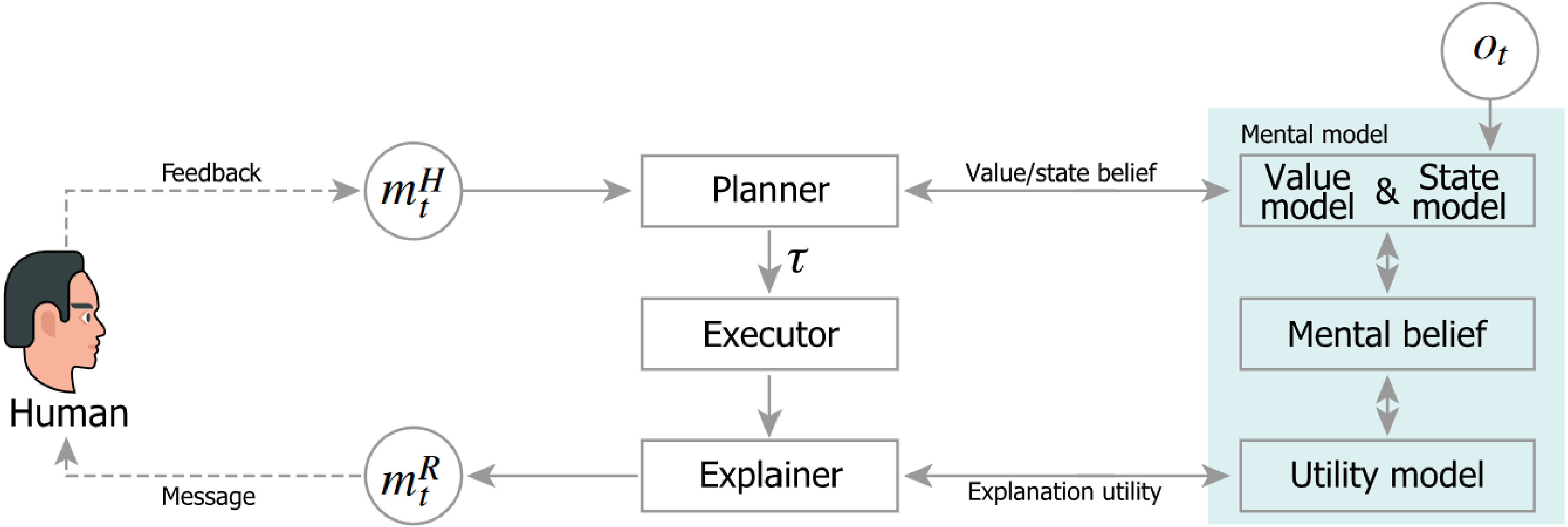




Step t
Step t+1
Processed state
Proposals & explanations
User feedback
$V^t_{UinM}$
$\delta^t_M$
$\delta^t_U$
$V^t_{MinU}$
$V_U$
$\delta^{t+1}_M$
$V^{t+1}_{UinM}$
$\delta^{t+1}_U$
$V^{t+1}_{MinU}$
Safety
Money
Time
Resource
Communicate
M to U
U to M
Estimate
M to U
U to M

# Agent Representation





Feedback
$m_t^H$
Planner
Value/state belief
Mental model
$o_t$
Value model
&
State model
$\tau$
Executor
Mental belief
Human
Explainer
Utility model
Message
$m_t^R$
Explanation utility

# Value Function

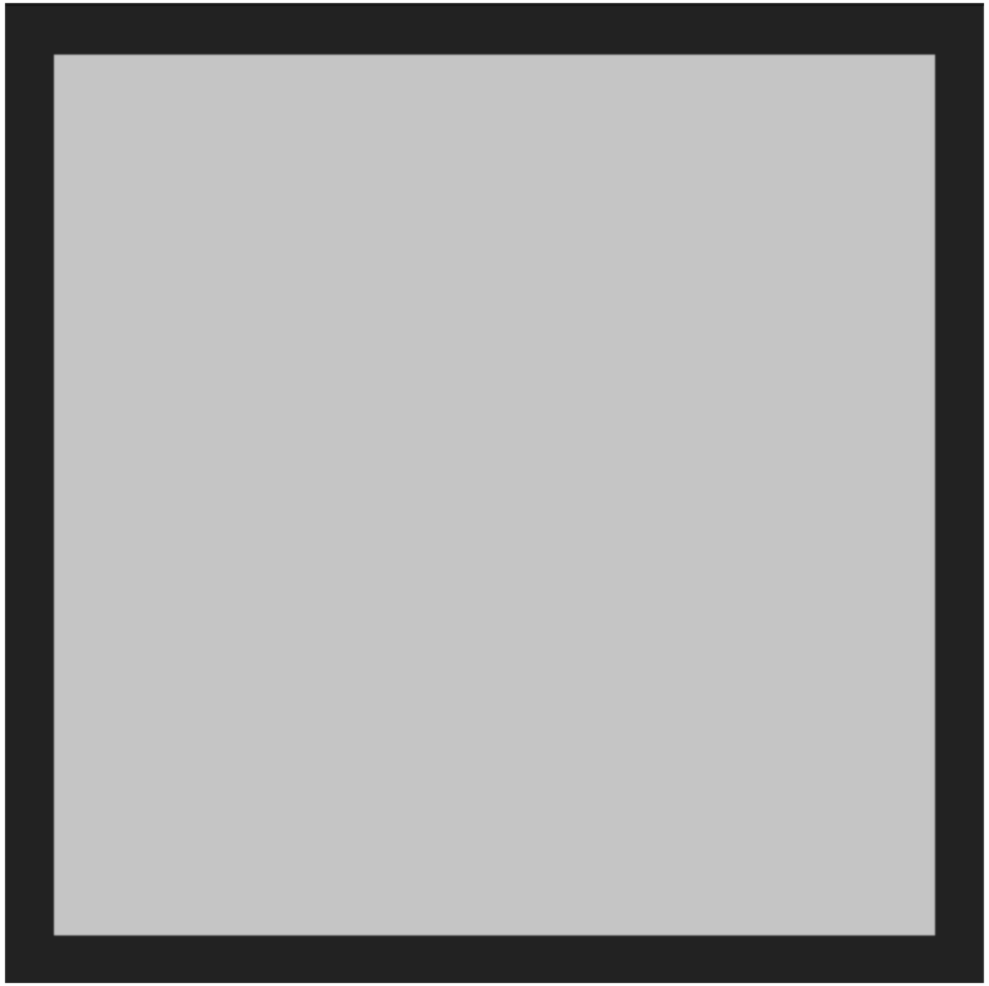
The importance of goals are modeled as a value function:

- Given robots action sequence, the task has certain measurements, each corresponds to a goal:
  - Total time used $\phi_T$
  - Number of resources collected $\phi_R$
  - Number of bomb defused $\phi_B$
  - Number of grids detected $\phi_D$
  - …. $\phi_i$
- The performance of the task is a value defined by the importance of each goal
  - The more important a goal is, larger the corresponding dimension of $\theta$ is

$$< \theta^T \phi > \quad = \quad < \theta_T, \phi_T > + < \theta_R, \phi_R > + < \theta_B, \phi_B > + < \theta_D, \phi_D > + \ \dots$$

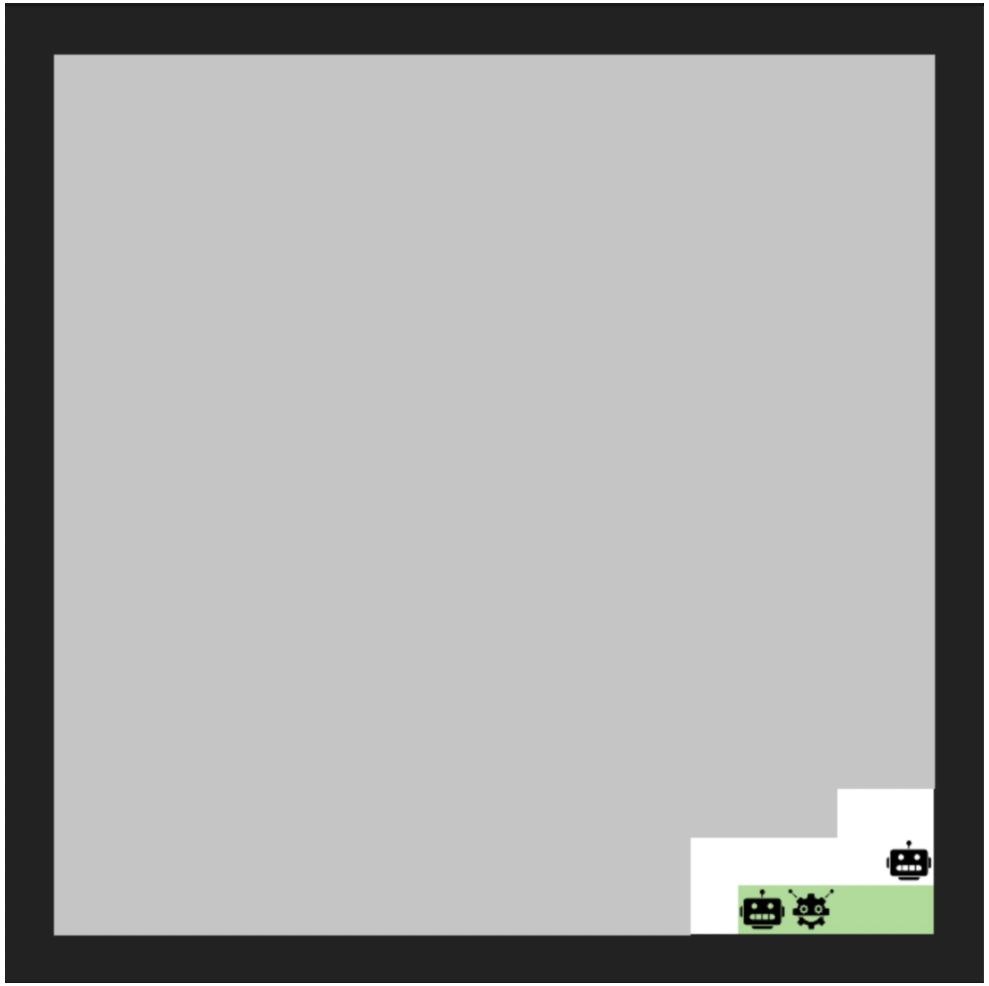
$$\|\theta\|_1 = 1$$

# Game Engine Progress



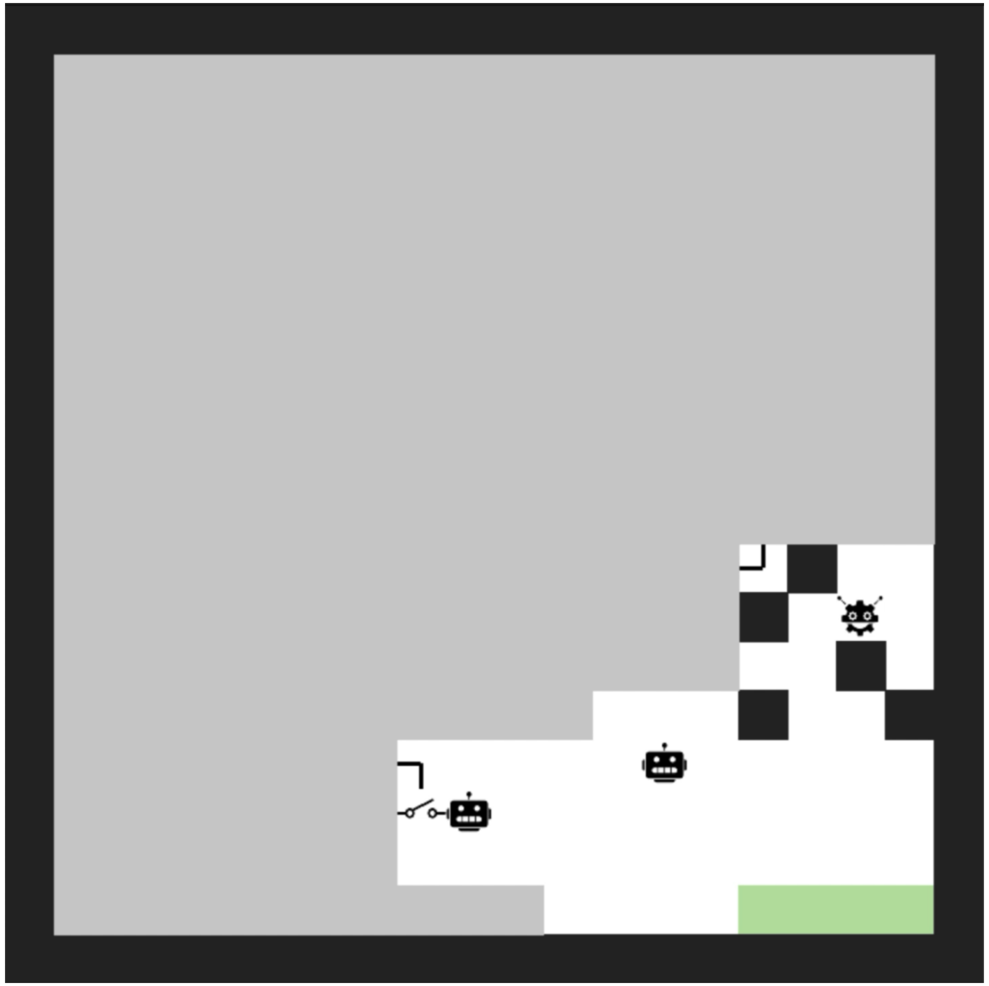
Scouts initialized

# Game Engine Progress



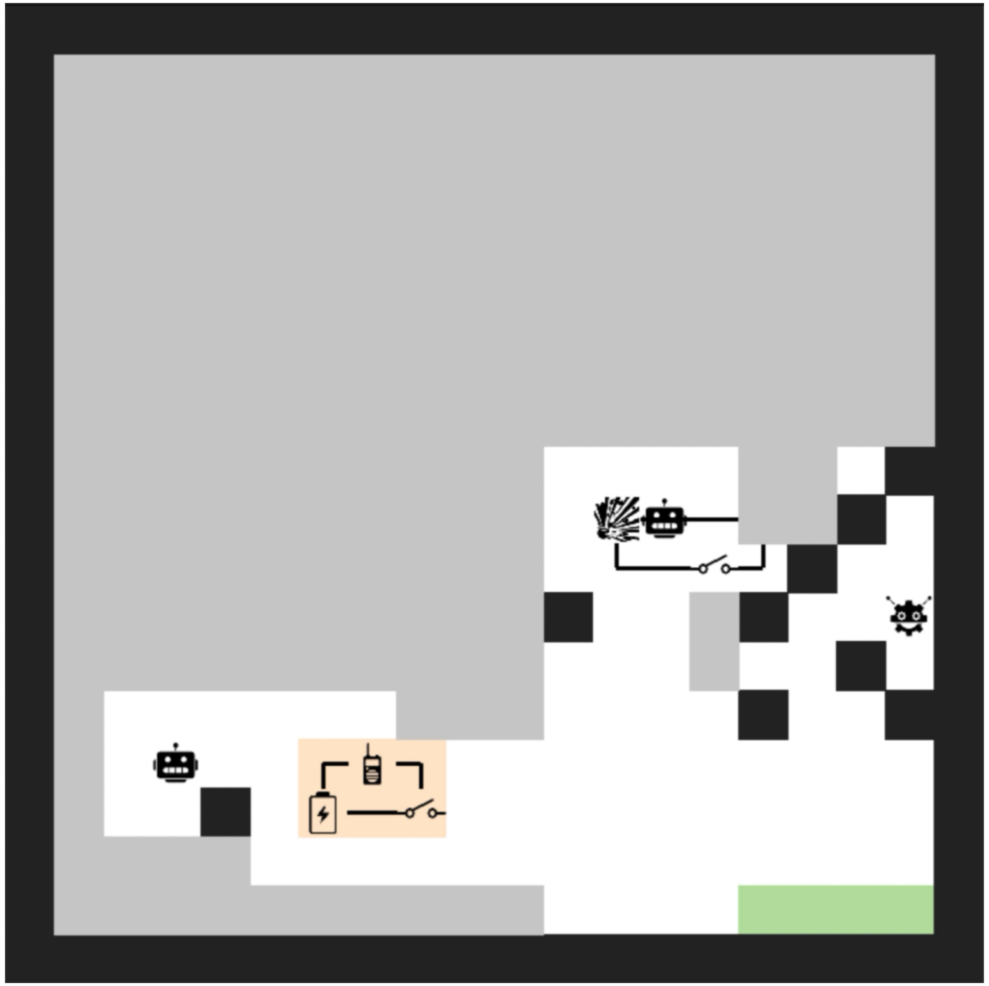
Scouts begin searching area

# Game Engine Progress



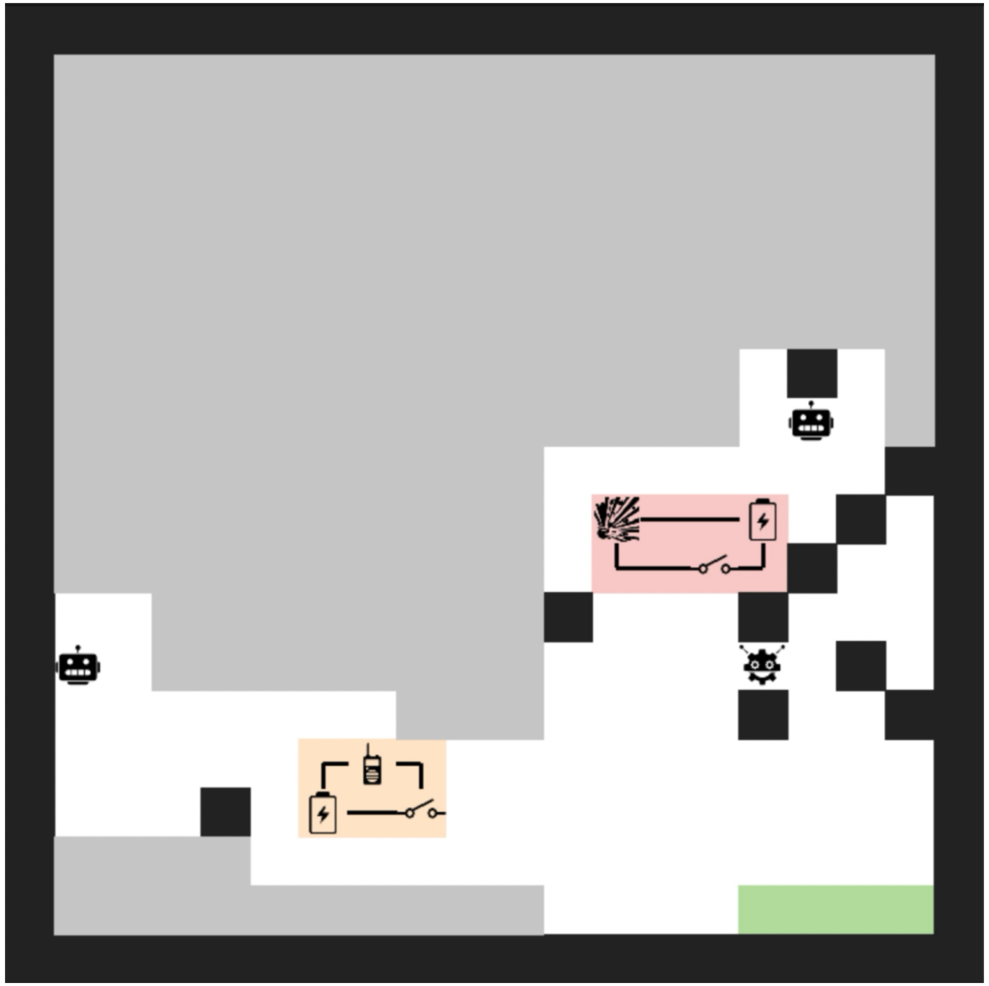
Suspicious device detected

# Game Engine Progress



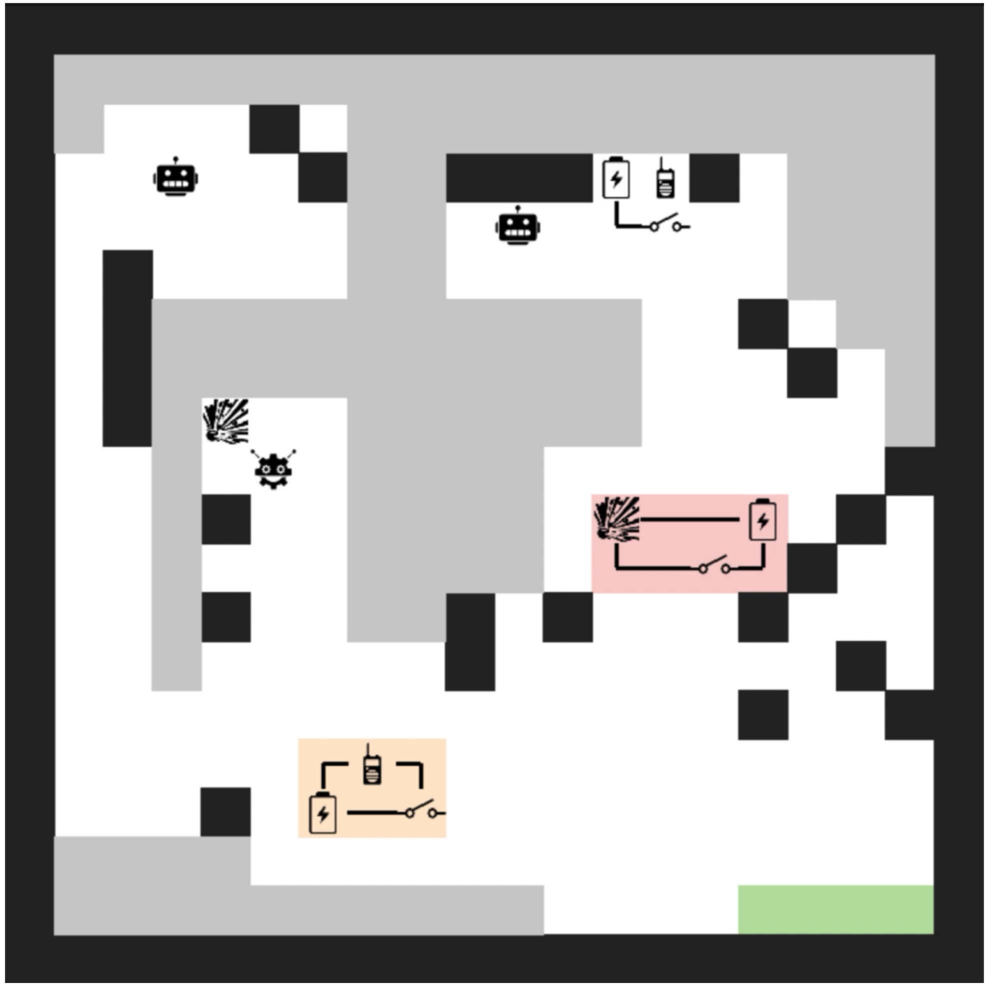
Bomb detected

# Game Engine Progress



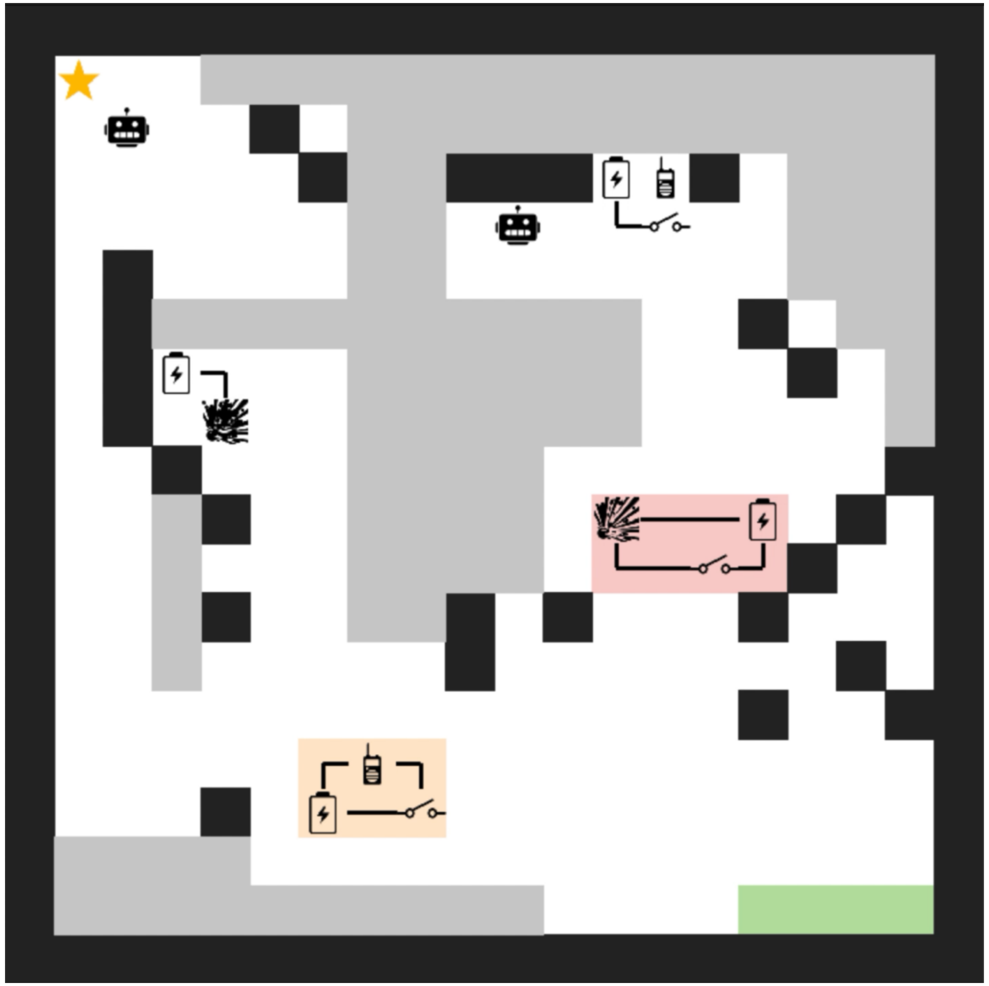
Exploration continues

# Game Engine Progress



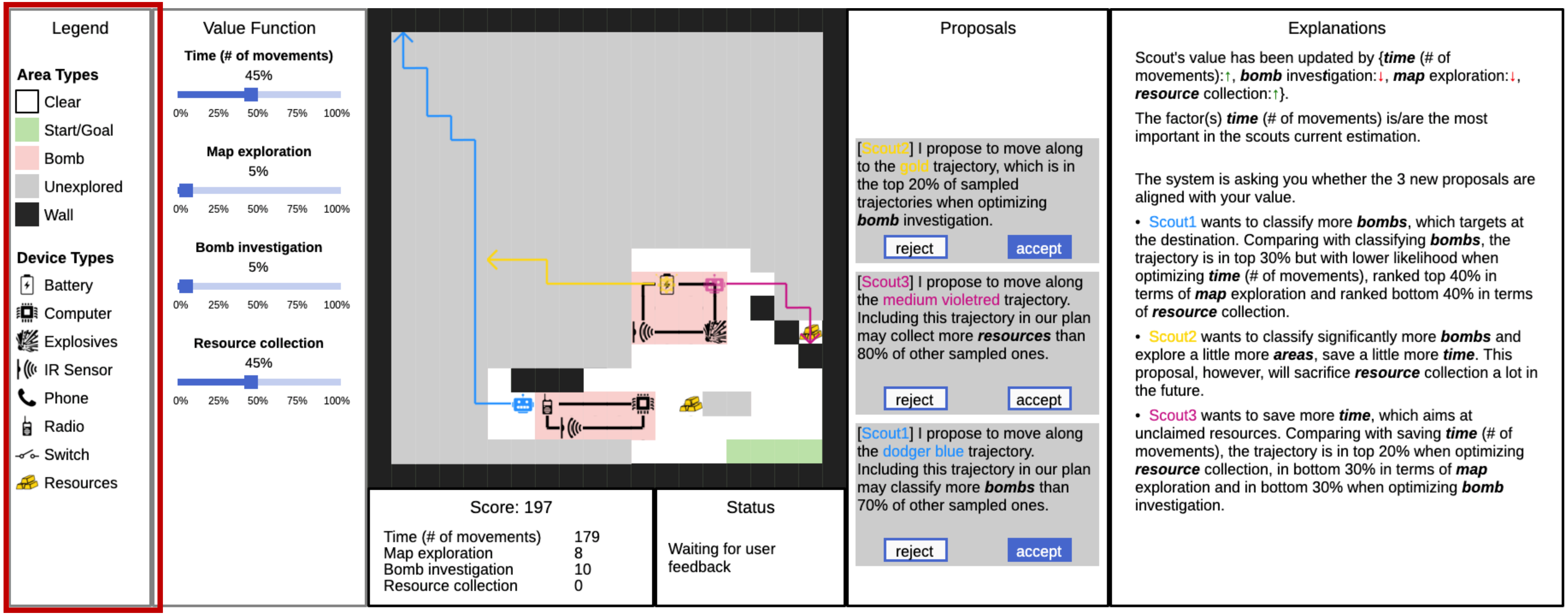
Goal discovered

# Game Engine Progress

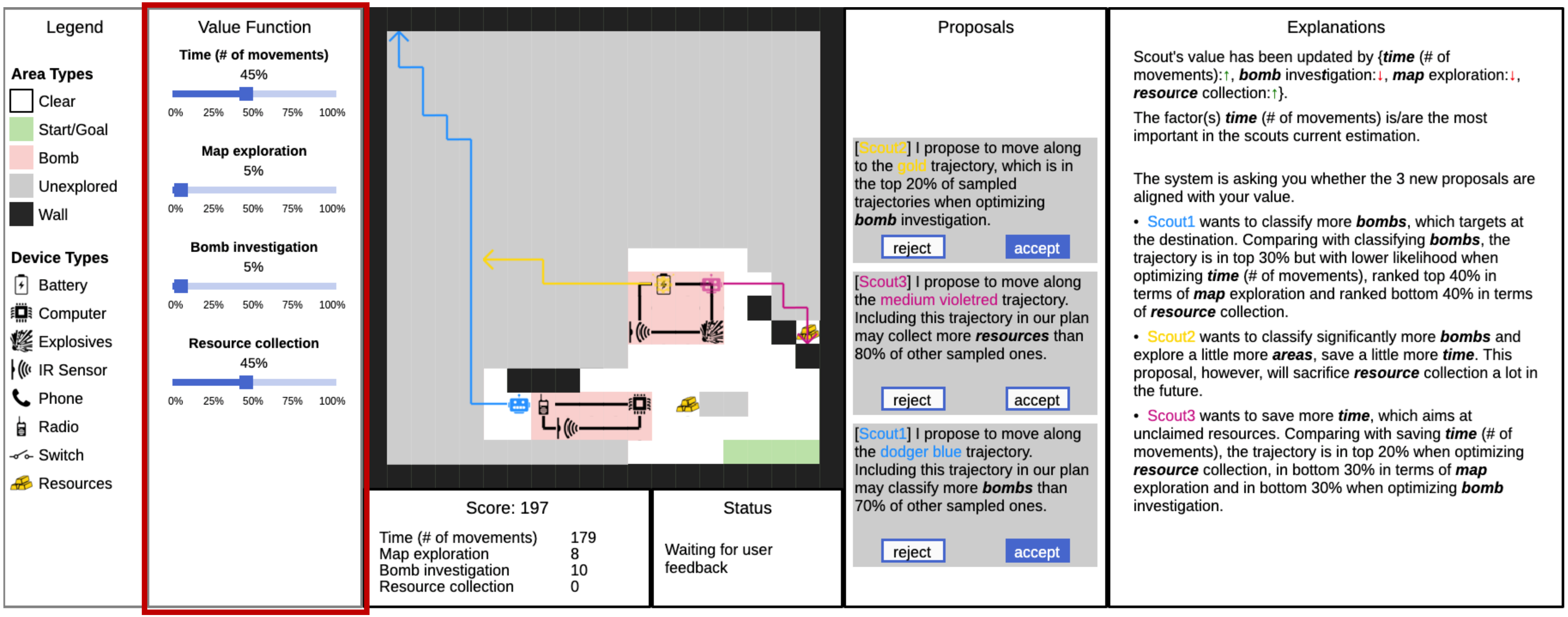


Second bomb detected

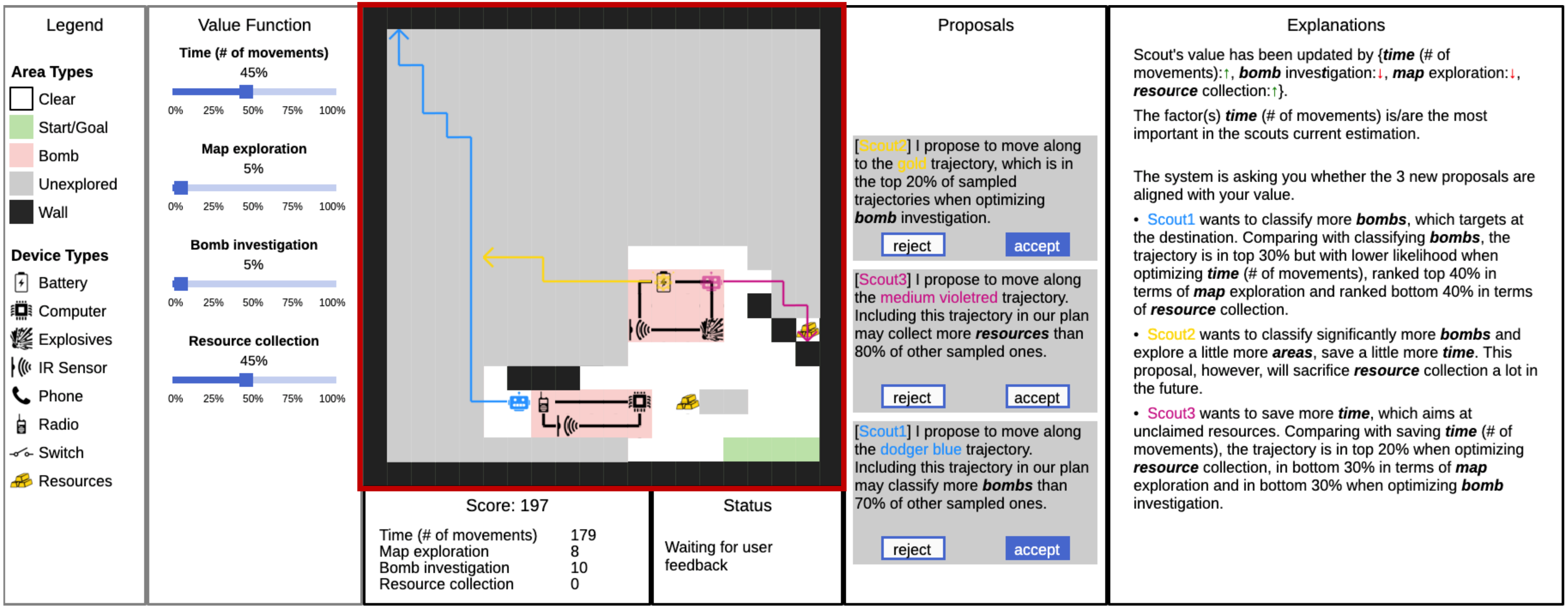
# Panel Introduction – Legend



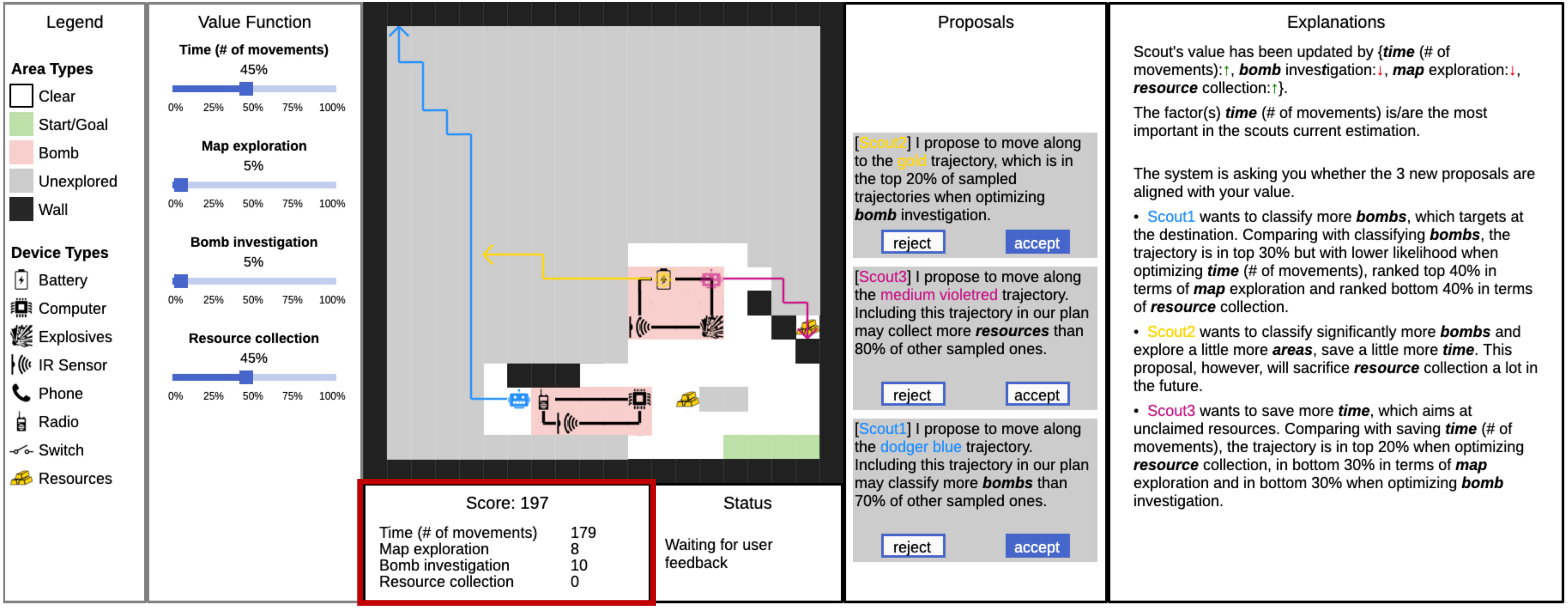
# Panel Introduction – Value Function



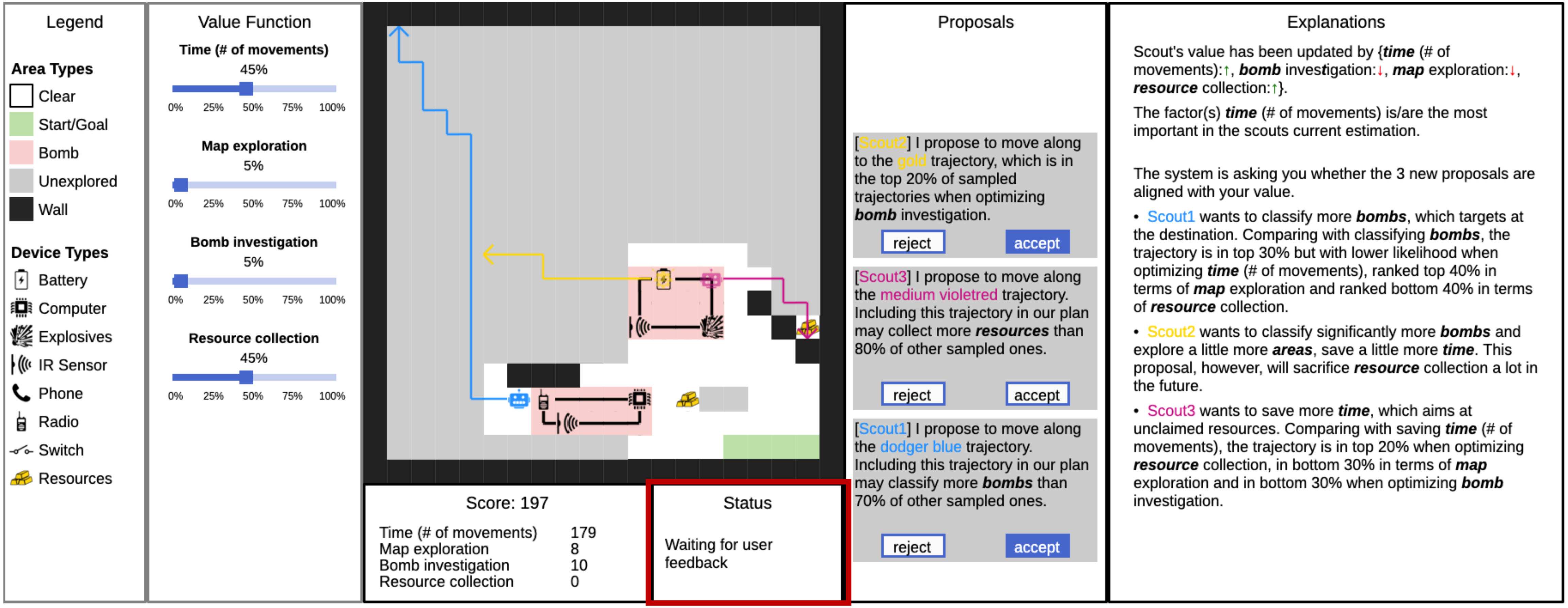
# Panel Introduction – Map



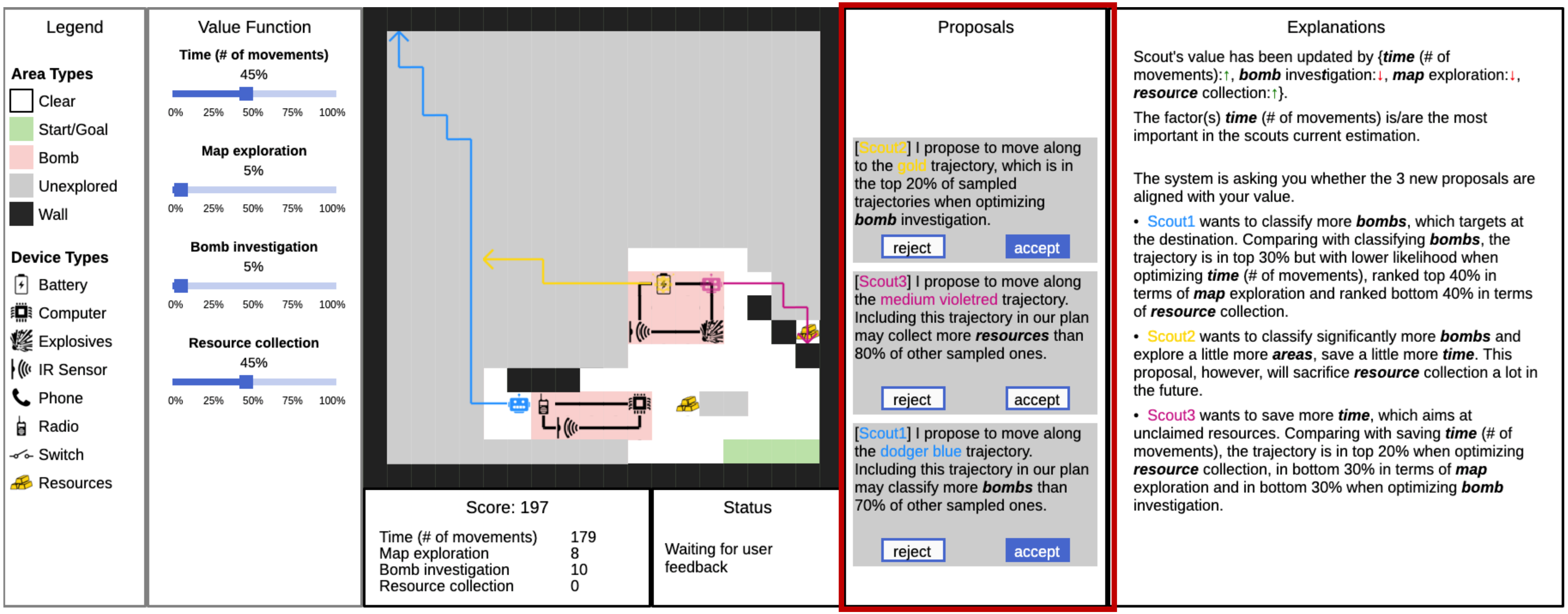
# Panel Introduction – Score



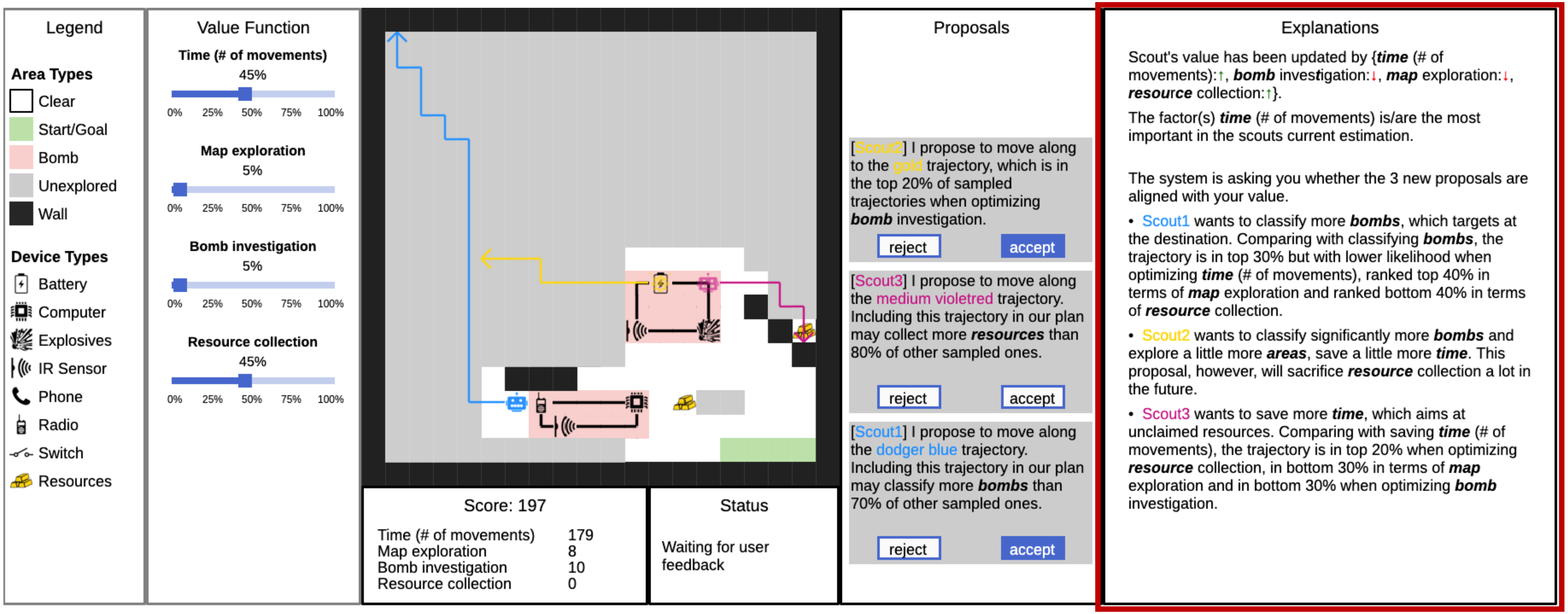
# Panel Introduction – Status



# Panel Introduction – Proposals



# Panel Introduction – Explanations



# Scout Exploration Game

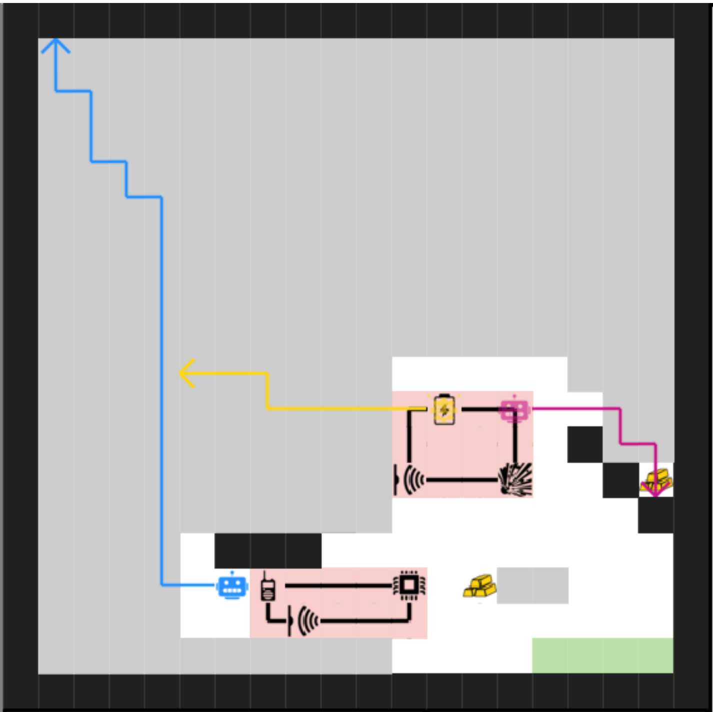
### Value Function

**Time (# of movements)**
45%
0% 25% 50% 75% 100%

**Map exploration**
5%
0% 25% 50% 75% 100%

**Bomb investigation**
5%
0% 25% 50% 75% 100%

**Resource collection**
45%
0% 25% 50% 75% 100%



### Score: 197

| | |
|---|---|
| Time (# of movements) | 179 |
| Map exploration | 8 |
| Bomb investigation | 10 |
| Resource collection | 0 |

### Status

Waiting for user feedback

### Proposals

[Scout2] I propose to move along to the gold trajectory, which is in the top 20% of sampled trajectories when optimizing ***bomb*** investigation.

reject accept

[Scout3] I propose to move along the medium violetred trajectory. Including this trajectory in our plan may collect more ***resources*** than 80% of other sampled ones.

reject accept

[Scout1] I propose to move along the dodger blue trajectory. Including this trajectory in our plan may classify more ***bombs*** than 70% of other sampled ones.

reject accept

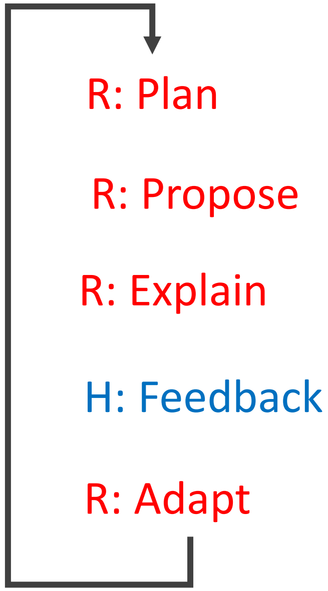
### Explanations

Scout's value has been updated by {***time*** (# of movements):↑, ***bomb*** investigation:↓, ***map*** exploration:↓, ***resource*** collection:↑}.

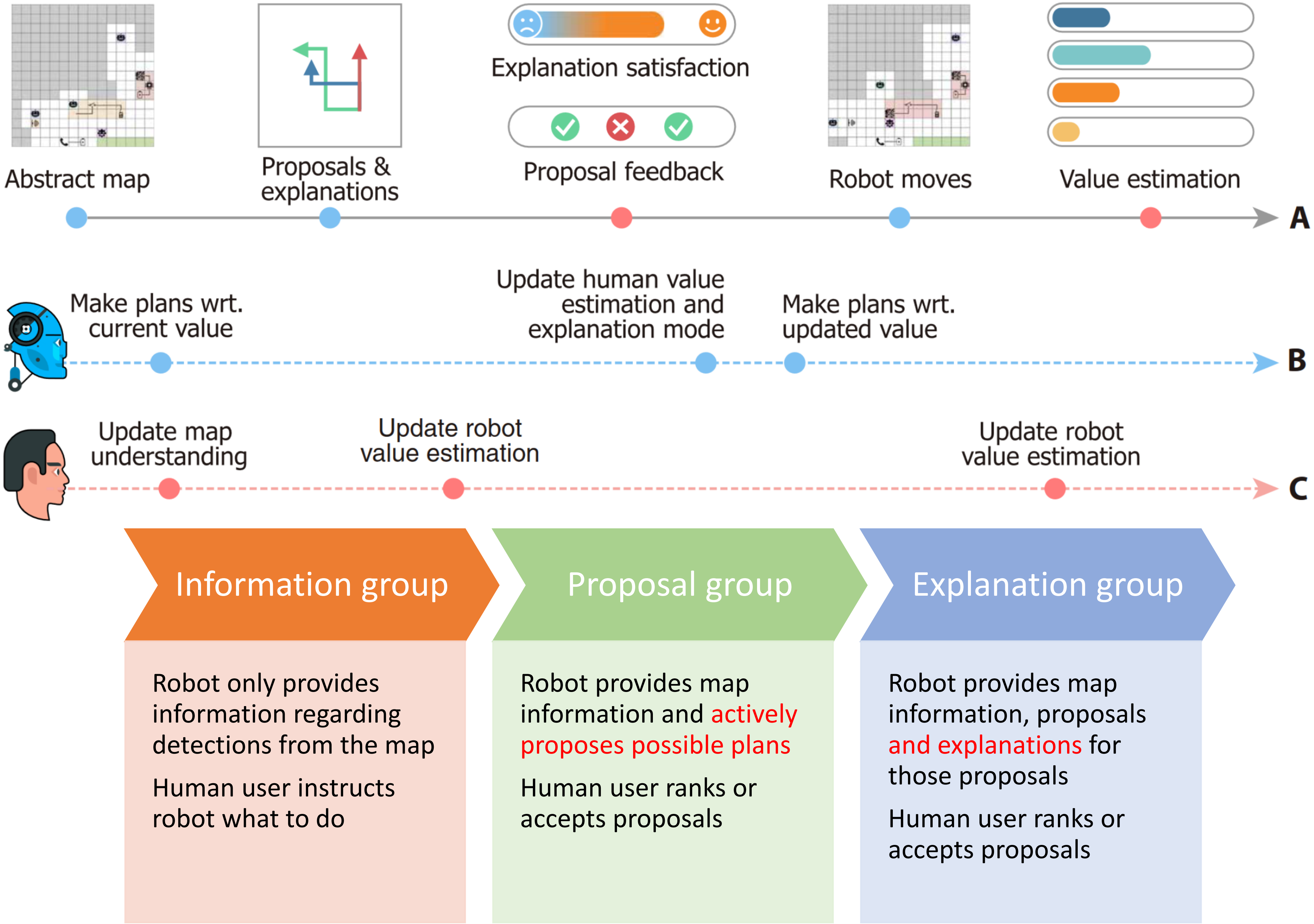
The factor(s) ***time*** (# of movements) is/are the most important in the scouts current estimation.

The system is asking you whether the 3 new proposals are aligned with your value.

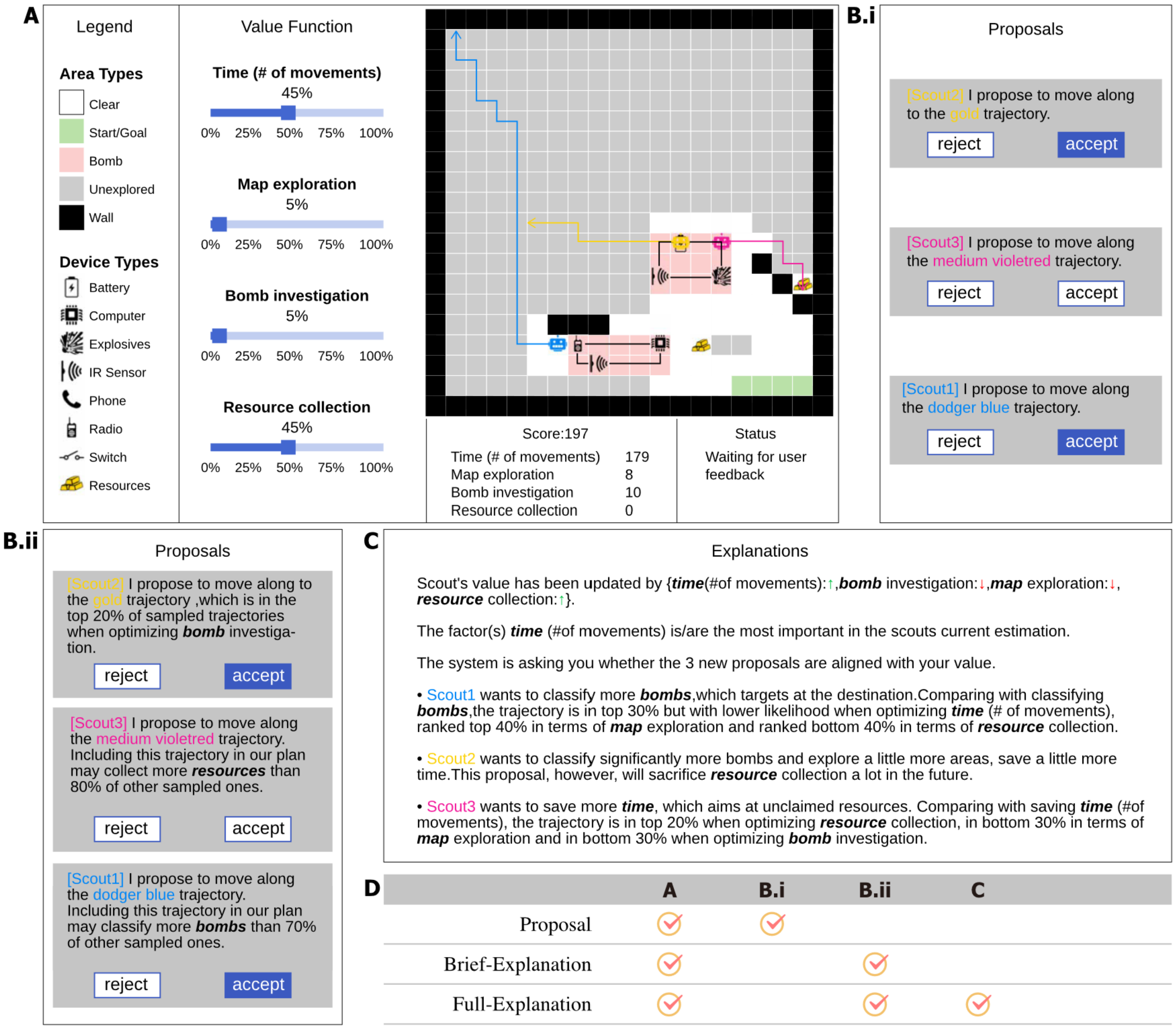
- Scout1 wants to classify more ***bombs***, which targets at the destination. Comparing with classifying ***bombs***, the trajectory is in top 30% but with lower likelihood when optimizing ***time*** (# of movements), ranked top 40% in terms of ***map*** exploration and ranked bottom 40% in terms of ***resource*** collection.
- Scout2 wants to classify significantly more ***bombs*** and explore a little more ***areas***, save a little more ***time***. This proposal, however, will sacrifice ***resource*** collection a lot in the future.
- Scout3 wants to save more ***time***, which aims at unclaimed resources. Comparing with saving ***time*** (# of movements), the trajectory is in top 20% when optimizing ***resource*** collection, in bottom 30% in terms of ***map*** exploration and in bottom 30% when optimizing ***bomb*** investigation.



Legend
Area Types
Clear
Start/Goal
Bomb
Unexplored
Wall
Device Types
Battery
Computer
Explosives
IR Sensor
Phone
Radio
Switch
Resources
Value Function
Time (# of movements)
5%
0% 25% 50% 75% 100%
Map exploration
5%
0% 25% 50% 75% 100%
Bomb investigation
5%
0% 25% 50% 75% 100%
Resource collection
85%
0% 25% 50% 75% 100%
Score: 54
Time (# of movements) 21
Map exploration 7
Bomb investigation 0
Resource collection 26
Status
Planning because circuit detection changed [0/100%]
Proposals
Explanations



Abstract map
Proposals & explanations
Explanation satisfaction
Proposal feedback
Robot moves
Value estimation
A
Make plans wrt. current value
Update human value estimation and explanation mode
Make plans wrt. updated value
B
Update map understanding
Update robot value estimation
Update robot value estimation
C
Information group
Robot only provides information regarding detections from the map
Human user instructs robot what to do
Proposal group
Robot provides map information and actively proposes possible plans
Human user ranks or accepts proposals
Explanation group
Robot provides map information, proposals and explanations for those proposals
Human user ranks or accepts proposals


A
Legend
Area Types
Clear
Start/Goal
Bomb
Unexplored
Wall
Device Types
Battery
Computer
Explosives
IR Sensor
Phone
Radio
Switch
Resources
Value Function
Time (# of movements)
45%
0% 25% 50% 75% 100%
Map exploration
5%
0% 25% 50% 75% 100%
Bomb investigation
5%
0% 25% 50% 75% 100%
Resource collection
45%
0% 25% 50% 75% 100%
Score:197
Time (# of movements) 179
Map exploration 8
Bomb investigation 10
Resource collection 0
Status
Waiting for user feedback
B.i
Proposals
[Scout2] I propose to move along to the gold trajectory.
reject
accept
[Scout3] I propose to move along the medium violetred trajectory.
reject
accept
[Scout1] I propose to move along the dodger blue trajectory.
reject
accept
B.ii
Proposals
[Scout2] I propose to move along to the gold trajectory ,which is in the top 20% of sampled trajectories when optimizing bomb investiga-tion.
reject
accept
[Scout3] I propose to move along the medium violetred trajectory. Including this trajectory in our plan may collect more resources than 80% of other sampled ones.
reject
accept
[Scout1] I propose to move along the dodger blue trajectory. Including this trajectory in our plan may classify more bombs than 70% of other sampled ones.
reject
accept
C
Explanations
Scout's value has been updated by {time(#of movements):↑,bomb investigation:↓,map exploration:↓, resource collection:↑}.
The factor(s) time (#of movements) is/are the most important in the scouts current estimation.
The system is asking you whether the 3 new proposals are aligned with your value.
• Scout1 wants to classify more bombs,which targets at the destination.Comparing with classifying bombs,the trajectory is in top 30% but with lower likelihood when optimizing time (# of movements), ranked top 40% in terms of map exploration and ranked bottom 40% in terms of resource collection.
• Scout2 wants to classify significantly more bombs and explore a little more areas, save a little more time.This proposal, however, will sacrifice resource collection a lot in the future.
• Scout3 wants to save more time, which aims at unclaimed resources. Comparing with saving time (#of movements), the trajectory is in top 20% when optimizing resource collection, in bottom 30% in terms of map exploration and in bottom 30% when optimizing bomb investigation.
D
A B.i B.ii C
Proposal
Brief-Explanation
Full-Explanation

# In Situ Bidirectional Human-Robot Value Alignment

Luyao Yuan[*†], Xiaofeng Gao[†], Zilong Zheng[†], Mark Edmonds[*], Ying Nian Wu, Federico Rossano, Hongjing Lu[*], Yixin Zhu[*], Song-Chun Zhu[*]

[*] Corresponding authors

[†] Authors contributed equally to this work

# Q & A

Thank you!