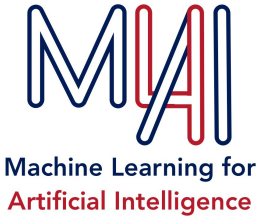


ToMCAT: Theory of Mind-based Cognitive Architecture for Teams

September 25, 2019

Bay of Pigs invasion (April 1961)

"All the News That's Fit to Print"

The New York Times.

LATE CITY EDITION

VOL. CX...No. 37,707. NEW YORK, THURSDAY, APRIL 20, 1961. FIVE CENTS

SENATE DEFEATS MONRONEY MOVE TO CURB PAY BILL

Passage Assured as Final Threat Is Beaten, 56-39, by Kennedy Forces

VOTE SCHEDULED TODAY

Goldwater Loses on Motion to Restrict Imports From Low-Wage Countries

By TOM WICKER

Goldberg Sets Up Office To Cope With Automation

Federal Unit, Which Opens Today, Will Study Manpower and Develop Plans to Retrain Displaced Workers

By STANLEY LEVEY

HOUSE VOTES BILL ON TO NEW JUDGES

Democrats Unite in 336-76

19 Police in Queens Shifted in Shake-Up Of Gambling Squad

By GUY PASSANT

U.S. TROOPS TO ACT AS FIELD ADVISERS TO LAOTIAN ARMY

300 to Drop Civilian Status —Britain Pressing Soviet for Cease-Fire Appeal

By JACQUES NEVARD

CASTRO SAYS ATTACK IS CRUSHED; CUBA REBELS GIVE UP BEACHHEAD, REPORT NEW LANDINGS ON ISLAND

REAR GUARD IS HIT

But Foes of Regime Tell of a Link-Up With Guerrillas

By TAD SZULC

BEFORE THE LANDINGS: Members of anti-Castro forces during maneuvers recently in the Caribbean area. The exiles took part in the various attacks on Cuba last Monday.

Cuban missile crisis (October 1962)

"All the News That's Fit to Print"

The New York Times.

LATE CITY EDITION

VOL. CXII...No. 38,258. NEW YORK, TUESDAY, OCTOBER 23, 1962. FIVE CENTS

U.S. IMPOSES ARMS BLOCKADE ON CUBA ON FINDING OFFENSIVE-MISSILE SITES; KENNEDY READY FOR SOVIET SHOWDOWN

U. S. JUDGES GIVEN POWER TO REQUIRE VOTE FOR NEGROES

High Court Upholds Order Forcing the Registration of 5A in Alabama County

By ANTHONY LEWIS

Chinese Open New Front; Use Tanks Against Indians

Nehru Warns of Peril to Independence —Reds Attack Near Burmese Border and Press Two Other Drives

SHIPS MUST STOP

Other Action Planned If Big Rockets Are Not Dismantled

By JAMES RESTON

ANNOUNCES HIS ACTION: President Kennedy speaking to the nation last night on radio and television. He told of moves to keep offensive equipment away from Cuba.

PRESIDENT GRAVE

Asserts Russians Lied and Put Hemisphere in Great Danger

By ANTHONY LEWIS

Planned by the *same team*!

The difference? Team dynamics.

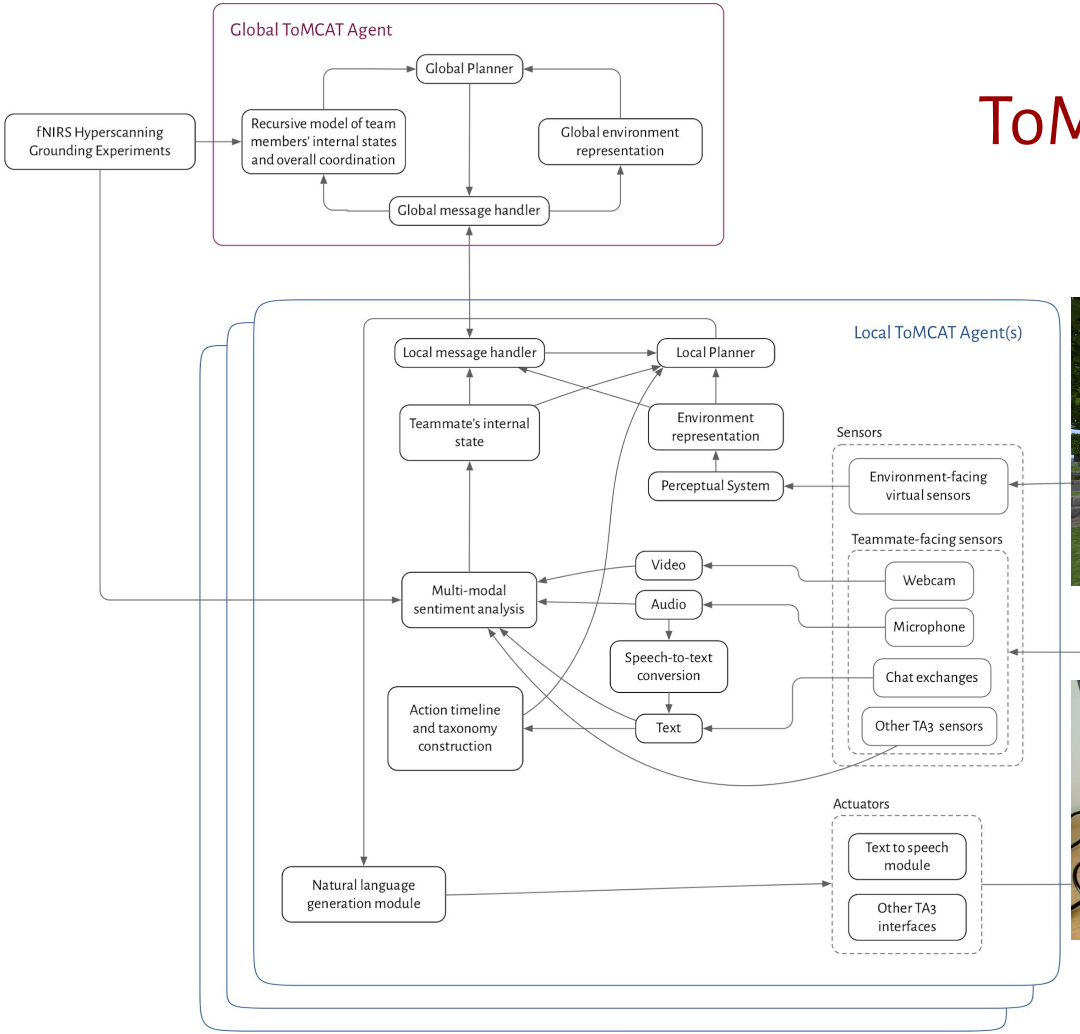
Source: Tetlock & Gardner, Superforecasting (2015)

What happens when one or more of the team members are machines?

Outline

- Architecture/overview & mission design (Adarsh)
- Hierarchical planning (Clay)
- Dialogue system (Mihai)
- Probabilistic Modeling (Kobus/Emily)
- fNIRS (Kobus/Emily)

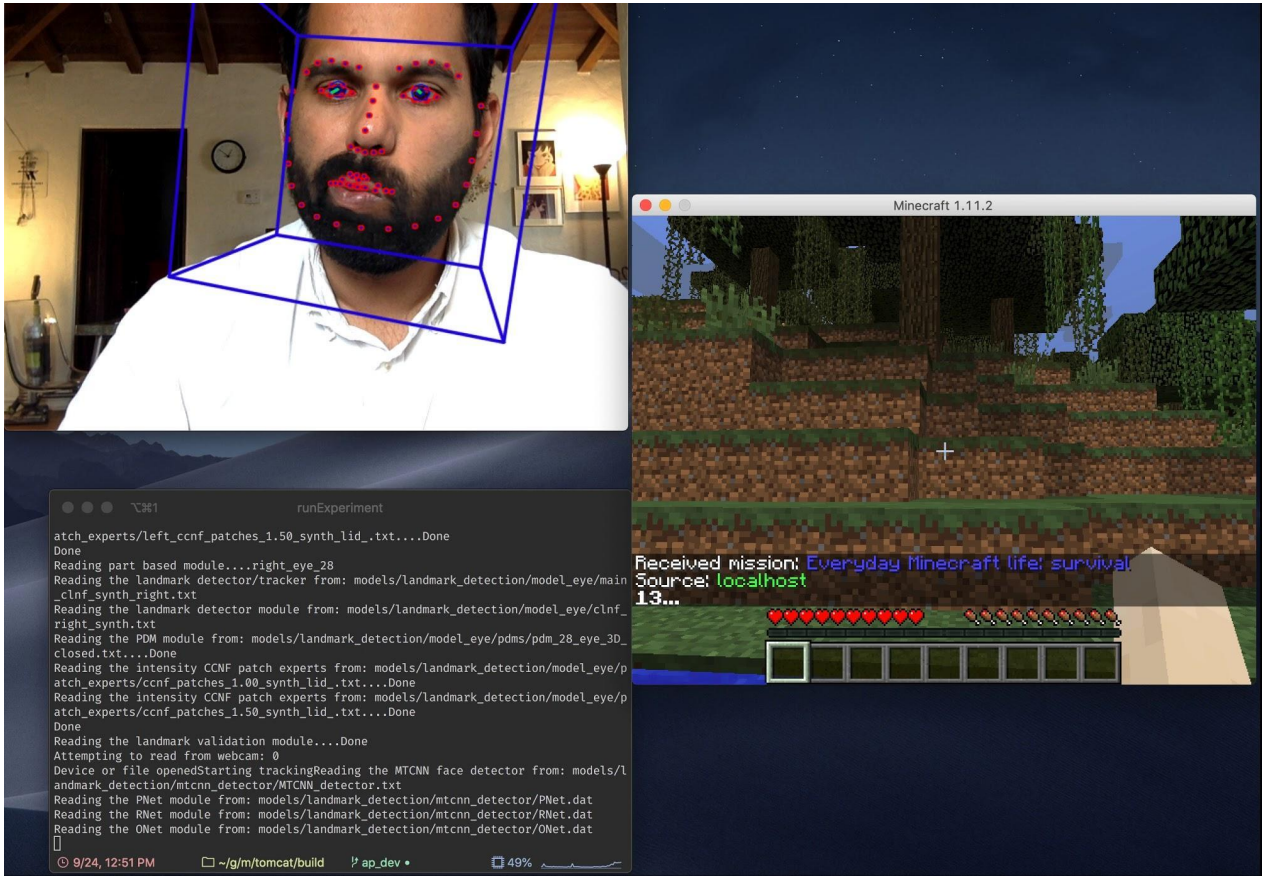
ToMCAT Architecture



Local environment

Human teammate

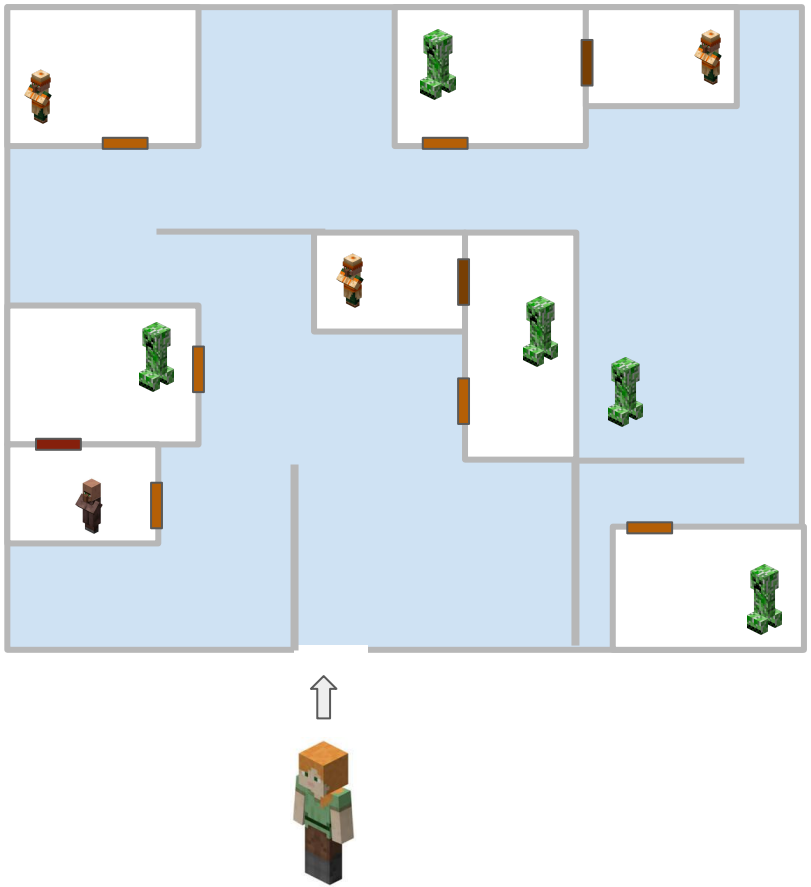
Webcam sensor integration



Mission design criteria

- Induces emotion (must not be boring!)
- Natural language processing and dialogue can play a significant role.
- Requires hierarchical planning and decision making (inherent in Minecraft's crafting system)
- Supports interesting multiplayer interactions (competition/cooperation)
- Supports perturbations

Starter mission: Search & Rescue with Crafting



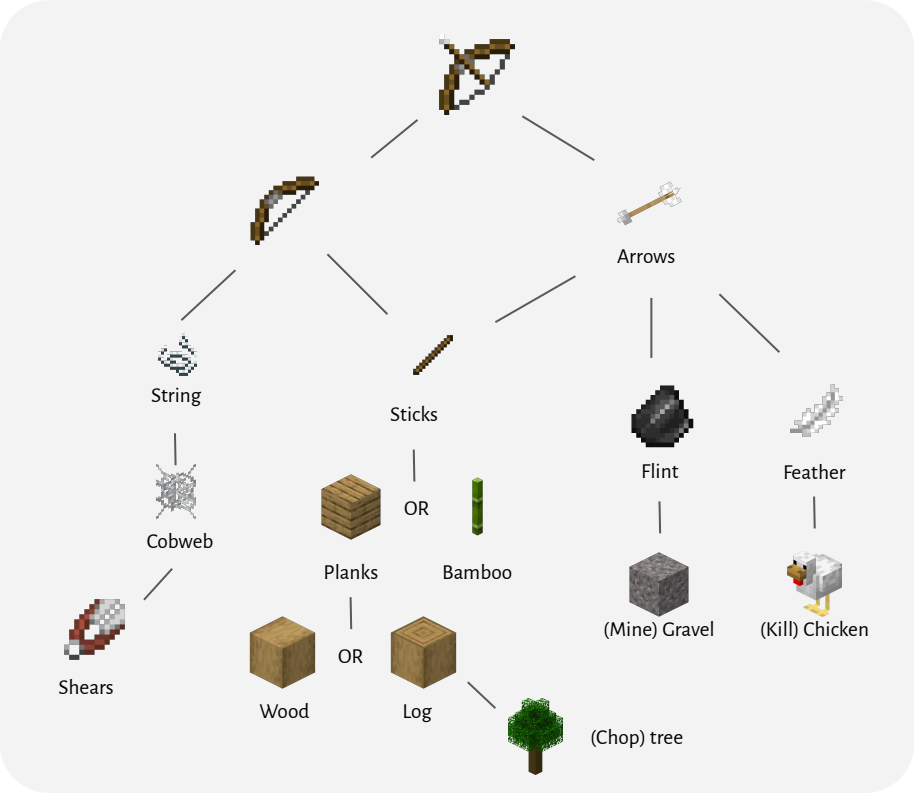
Context: Creepers have attacked the village, and the survivors are holed up in rooms in their houses, having blockaded the doors against the creepers. But they haven't got long…

Goal: Rescue villagers from creeper-infested village, within some time limit.

Features

- Induces emotion (time limit, creepers are dangerous)
- Hierarchical planning (need to craft a weapon to defend against creepers)
- Plan monitoring (need to monitor player health)
- Natural language dialogue (the player can get hints from the agent)
- Natural extension to multiplayer mission.

Starter mission: Search & Rescue with Crafting



Hierarchical planning - need to craft a weapon to defend ourselves from creepers!

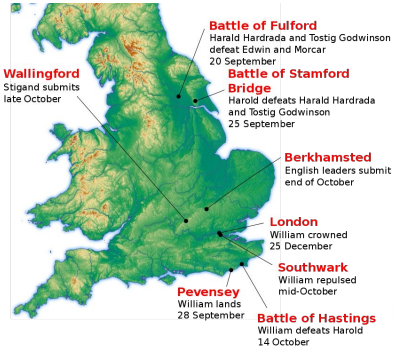
Near-term goals

- Refine missions
- Test internally within group while hooked up with basic autonomic physio
- Submit HSR application to UA IRB

ToMCAT Plan Representation

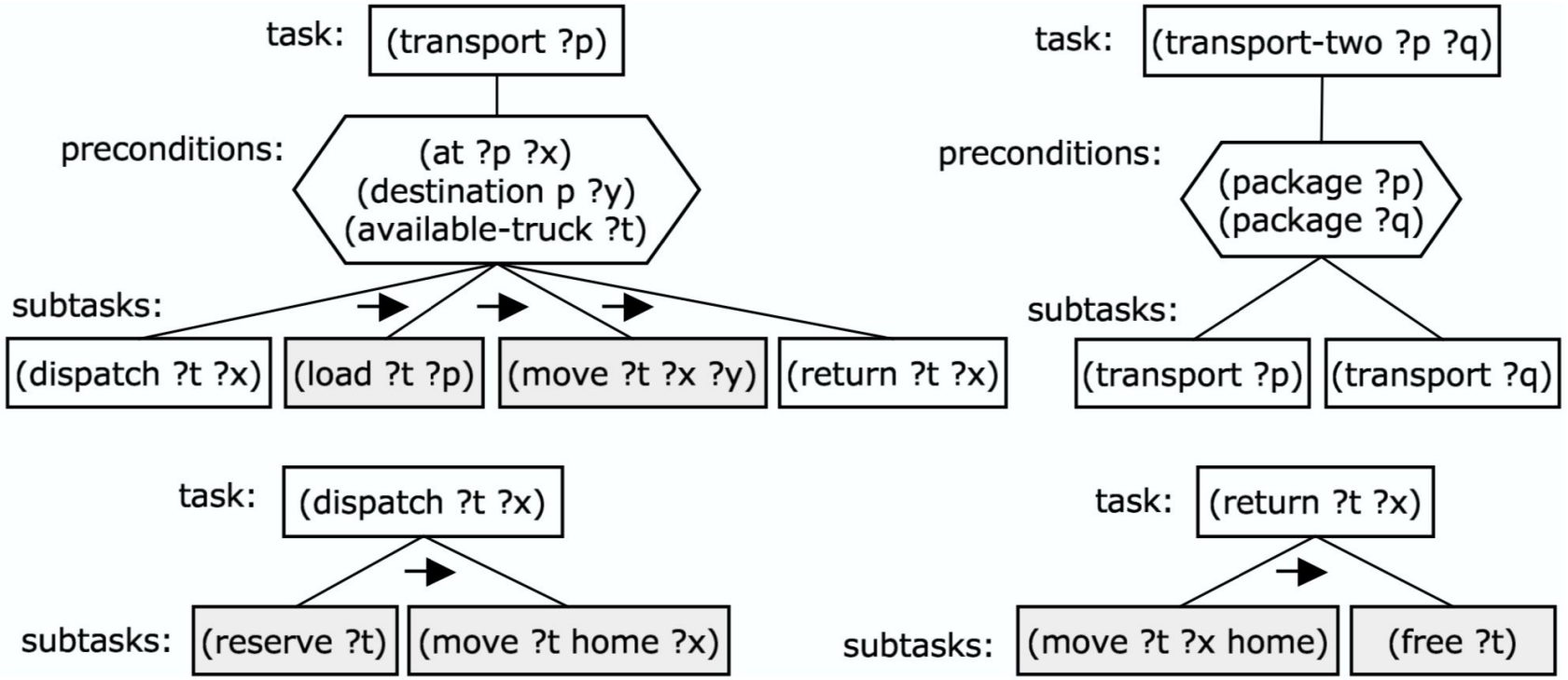
Hierarchical planning

- Understand what a plan is (plan representation)
  - Plan representation is hierarchical: different levels of abstraction / chunking
  - PDDL, HTN instance
- Track where the team is in plan execution
  - Plan recognition
  - Execution monitoring
- Update/modify plan
  - Planning: in face of missing information (assumption-based planning)
  - Plan repair
- Learn
  - Operators and methods
  - Start with Word2HTN

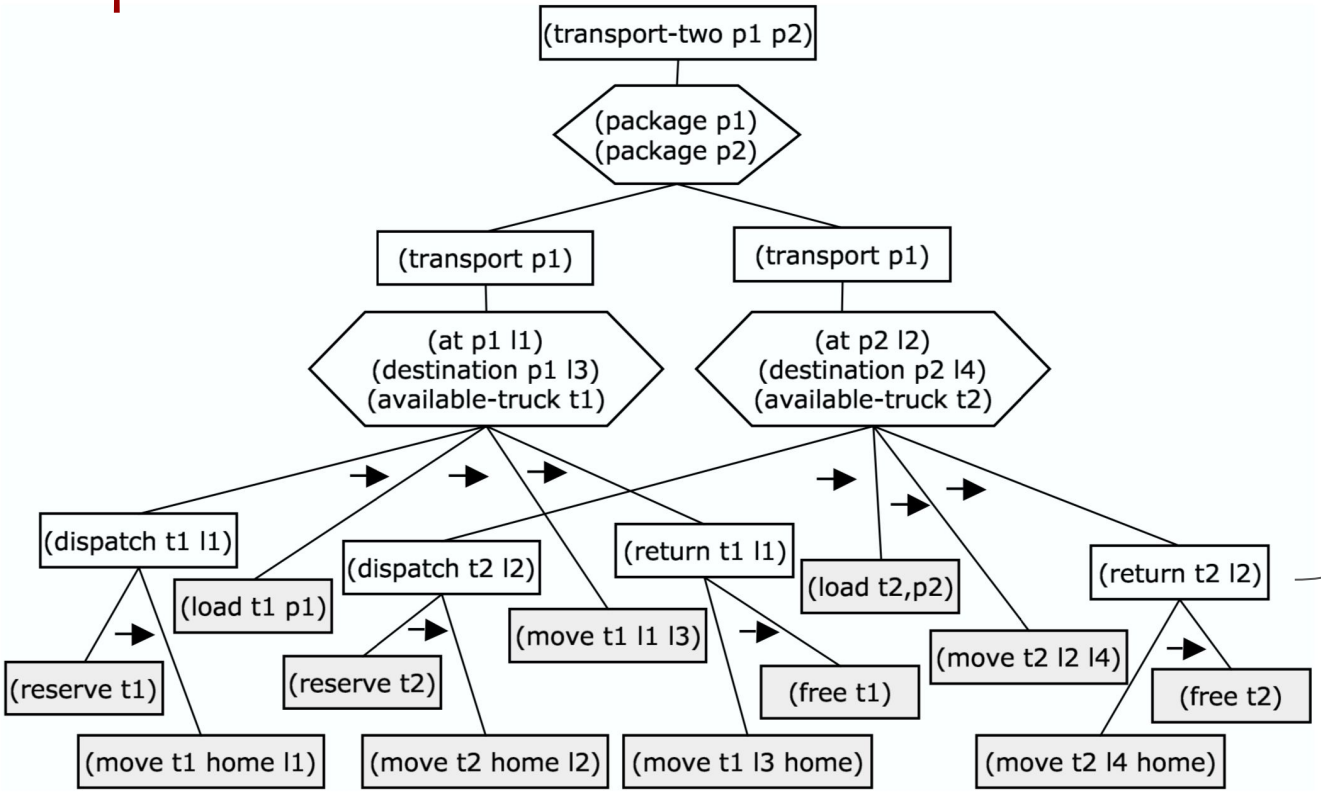


Source: *Robert P. Goldman and Ugur Kuter. 2019. Hierarchical Task Network Planning in Common Lisp: the case of SHOP3. In Proceedings of ELS '19: European Lisp Symposium (ELS '19). ACM, New York, NY, USA, 8 pages. https://doi.org/10.5281/zenodo.2633324*

Examples of Planning Building Blocks (operators, methods)



Example HTN Plan

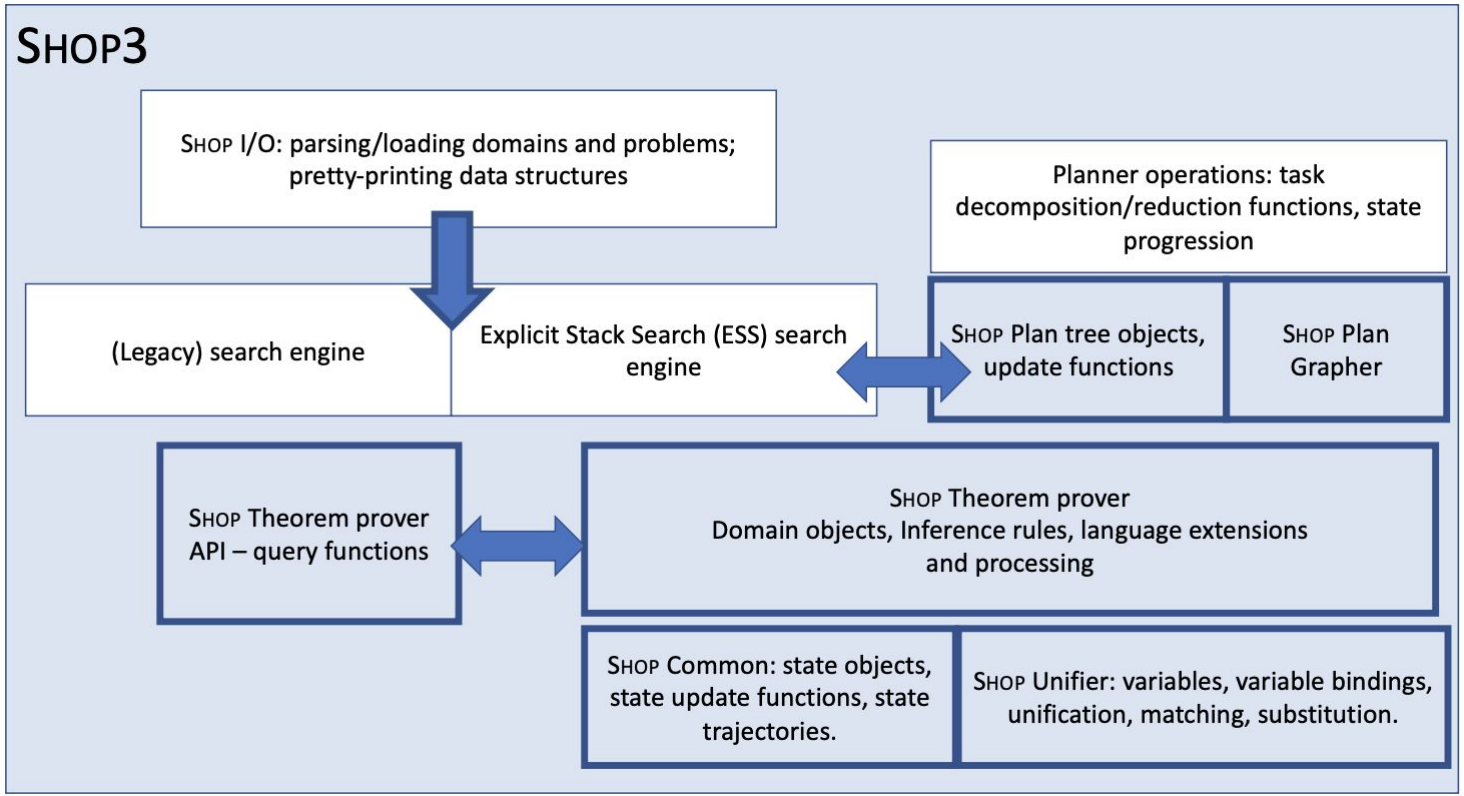


Main focus of ToMCAT:
- Locus of dialogue
- Coordination
- Plan repair / replanning

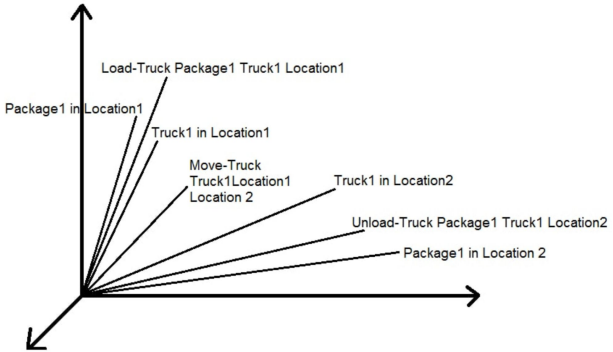
The "Plan"

Planning Domain Definition Language (PDDL)

- Basic STRIPS-style actions
- Conditional effects
- Universal quantification over dynamic universes (i.e., object creation & destruction)
- Domain axioms
- Specification of safety constraints
- Specification of hierarchical actions composed of subactions and subgoals
- Management of multiple problems in multiple domains using different subsets of language features (various levels of expressiveness)


SHOP3
SHOP I/O: parsing/loading domains and problems; pretty-printing data structures
Planner operations: task decomposition/reduction functions, state progression
(Legacy) search engine
Explicit Stack Search (ESS) search engine
SHOP Plan tree objects, update functions
SHOP Plan Grapher
SHOP Theorem prover API – query functions
SHOP Theorem prover
Domain objects, Inference rules, language extensions and processing
SHOP Common: state objects, state update functions, state trajectories.
SHOP Unifier: variables, variable bindings, unification, matching, substitution.

Learning HTNs: Word2HTN



Preconditions `package1 in Location1, Truck1 in Location1`
Action `Load package1 Truck1 Location1`
Effects `package1 in Truck1, Truck1 in Location1`

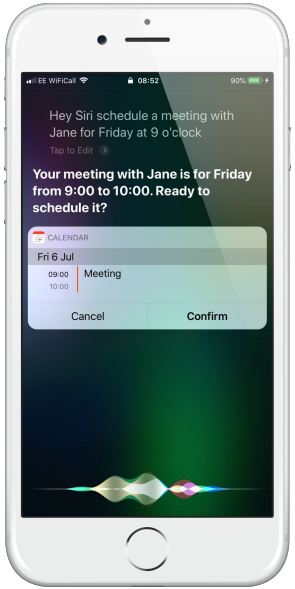
Preconditions `Truck1 in Location1, Truck1 canReach Location2`
Action `Move Truck1 Location1 Location2`
Effects `Truck1 in Location2`

Preconditions `package1 in Truck1, Truck1 in Location2`
Action `Unload package1 Truck1 Location2`
Effects `package1 in Location2, Truck1 in Location2`

ToMCAT Dialogue System

Intuition

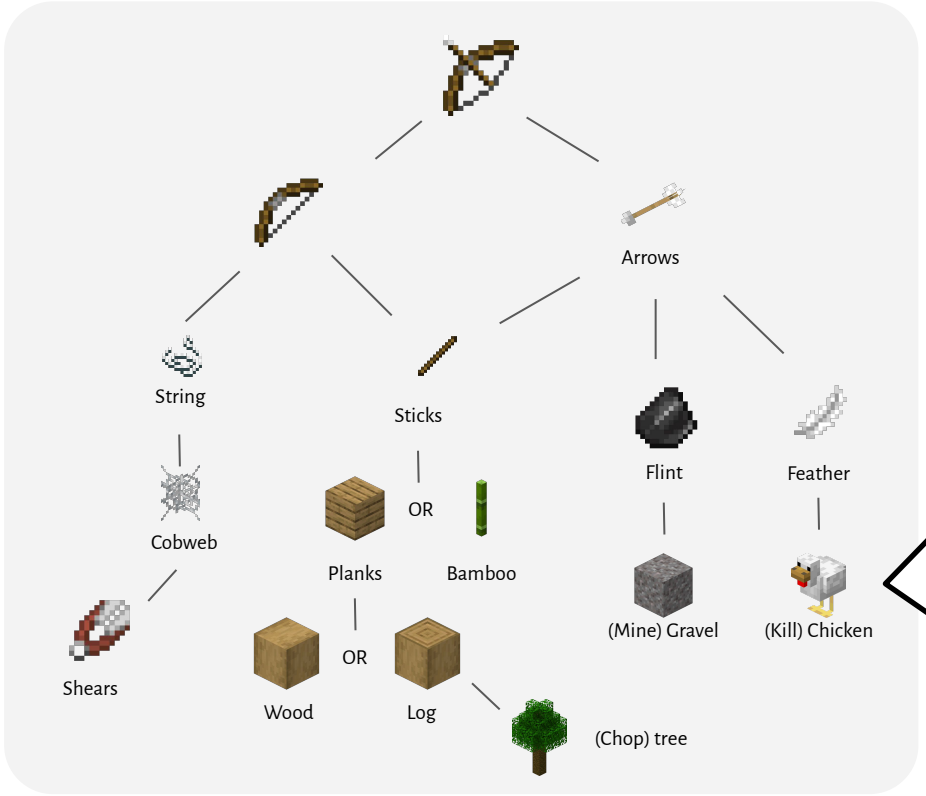
Task-based



+

Socially-aware

Action and timeline extraction: short term



- Each one of these actions will be mapped into an Odin grammar, e.g., “I’m cutting a tree”→ Chop tree
- Odin/Odinson: robust to ungrammatical text but can use semantic/syntactic representation when available

Action and timeline extraction: longer term

- Hybrid reading for action extraction: combining rules with neural methods
  - Use rule extractions as training data for neural extractors
  - Joint learning (see WM effort)

- Improve static action taxonomies with taxonomies that are dynamically extracted from game logs
  - See WM effort
  - Bad: explicit patterns (Hearst) infrequent, poor syntax
  - Good: enough data in gaming logs to generate (some) training data for neural methods; enough data to build/fine-tune distributed representations for implicit information

Action and timeline extraction: longer term

- Timeline extraction
  - Informed by linguistic patterns (see our Big Mechanism work)
    - Rules must be customized for gaming language rather than the formal language used in publications
      - Not too bad, less than 40 rules
  - Maybe: hybrid approach, combining rules with neural methods trained on the dataset of Hahn-Powell et al. (2016)

Timeline extraction: Rules vs. NNs (from Hahn-Powell, 2018)

	Model	*P* (%)	*R* (%)	*F1* (%)
Rules	Causal patterns	**85**	56	68
	Reichenbach	0	0	0
Feature-based	LR+L1	79	76	78
	LR+L2	79	75	77
	SVM+L1	80	78	**79**
	SVM+L2	77	74	75
	RF	77	80	78
Latent	LSTM	75	73	74
	LSTM+P	77	75	76
	FLSTM	78	76	77
	FLSTM+P	**83**	74	78
Sieves	Combined (all)	67	**90**	77

Multimodal sentiment analysis: short term

- Include an ASR in the ToMCAT system
  - Before sentiment analysis: can we extract actions from speech? How does this impact the corresponding Odin grammars?
    - Odin supports semantic match (based on embedding similarity); adding support based on phonetic distance would not be too hard

Multimodal sentiment analysis: longer term

- Multimodal analysis
  - Speech cues (e.g., pitch, vowel duration, burst amplitude, etc…)
  - Textual cues (plenty in gaming!)
  - Facial cues: smiling, frowning
  - Gaze
- We will start by using an existing open-source, multimodal framework for sentiment analysis (Zadeh et al., 2018)
  - Need to adapt contextualized embeddings from open domain to the gaming domain
  - Add previous utterances from the current dialogue

Cross-sentence coreference resolution

- Need to understand the *object* of actions (“chopping *it*”) and sentiment (“I love *it*”)
- Generic coreference resolution methods do not work on texts that do not come from the news domain (Bell et al., 2016)
- Data will be scarce here…
  - Cannot really train a big neural model

Cross-sentence coreference resolution

- Sieve-based architecture
  - Groups of rules applied in descending order of precision
  - A lower-precision rule cannot override a higher-precision one
  - Incrementally and cautiously builds a global representation of the concepts discussed
- We plan to mix open domain sieves with game-specific ones (see Bell et al., 2016)
  - Some sieves may use ML (see Lee et al., 2017)
- We plan to take advantage of the game context
  - “Incoming at 3 o’clock!”

Dialogue agent

- Monitors adherence to the plan using the components discussed previously
- Participant not enforcer!
  - Generate language to suggest ideas using a template-based language generator approach
  - Jointly maximize the task goal and the social goal
  - Beam search over the top *k* plans produced by ToMCAT
    - Back off to a lower-ranked plan that might increase the engagement of the human participants

ToMCAT Team ToM

Representing, measuring, and inferring what teaming minds are up to

Prepared by Emily Butler and Kobus Barnard

On behalf of our rapidly growing sub-team:

Emily Butler, Kobus Barnard, Adarsh Pyarelal, Clayton Morrison, Paulo Soares, Savannah Boyd, Ashley Kuelz, Harry Go, Lize Chen*, Jianfeng Li*, and Aditya Banerjee*

*Undergraduates helping build Minecraft missions that induce affect.

Initial activities and plans for getting started

- Mission modeling
  - Get simple missions into the lab ASAP to see what works for ease of use and **physiological induction**
  - In parallel with simple dialogue and planning
  - Starting with single person game
- Team theory literature review
  - Needed to better anticipate TA2 needs
  - Underway by Emily and her students Savannah Boyd and Ashley Kuelz
- Refined equipment plan

Why probabilistic graphical modeling of team minds?

- Theories matter for ASIST
  - Need representation for theoretical constructs
- We expect ongoing reconfiguring and extending of models
- Probability estimates support risk calculations (e.g., triage scenario)

Why dynamical Bayes nets for the team minds PGM?

- Intervention matters for ASIST
  - Need to be able to simulate what will happen if something is changed
- Some (most?) theories will entail a causal story
  - Thus we need conditioning to be forward in time
    - Analogous to ‘unrolling’ CAGs in WM

ToMCAT's modeling of the peoples' minds

- TomCAT will model affect and beliefs for each person

- Affect breaks down into
  - Overall affect
  - Affect about someone else
  - Affect about the team as a whole
  - Affect towards ToMCAT
  - Affect about the task or environment.

- Belief of a person, *P*, breaks down into
  - Beliefs about the task and the environment
  - Beliefs about others affect towards them
  - Beliefs about others affect towards other individuals, the team, and ToMCAT

Additional interesting complexities

- Some observables (e.g., speech with intonation) will be broken into affect (e.g., anger) and semantics (about what, and towards whom).

- ToMCAT observation model includes knowing what data each person has access to
    - For example, via eye tracking, ToMCAT may know that person A did not see person B's face at a critical time

Notation (purple colored symbols appear in the Bayes Net)

Superscript (t) $\Leftrightarrow$ time t

N $\Leftrightarrow$ number of persons

N_* $\Leftrightarrow$ number of measures of action (N_A), individual expression behavior (N_B), social behavior (N_S), coordination (N_C), and performance (N_V)

i $\Leftrightarrow$ indexes persons and team as a whole

j $\Leftrightarrow$ indexes human partners, team, and ToMCAT

m $\Leftrightarrow$ indexes measures (observation modalities)

External deterministic parameters

P $\Leftrightarrow$ Plans

E $\Leftrightarrow$ Environment (excluding L)

L_{ijm} $\Leftrightarrow$ Physical communication link

G $\Leftrightarrow$ Team groups (sub-teams)

Social parameters

T_i $\Leftrightarrow$ ToMCAT model of beliefs and affect of person i

T_{ij} $\Leftrightarrow$ ToMCAT model for person i affect towards j

O_{ijm} $\Leftrightarrow$ Person i observes person j with measure m

C $\Leftrightarrow$ ToMCAT model of coordination

Transition hyperparameters for social parameters
(stationary, some individual aspects omitted)

$\Theta^T \Leftrightarrow T_i$ $\quad \Theta^{T2} \Leftrightarrow T_{ij}$ $\quad \Theta^O \Leftrightarrow O$ $\quad \Theta^C \Leftrightarrow C$

Observations (evidence)

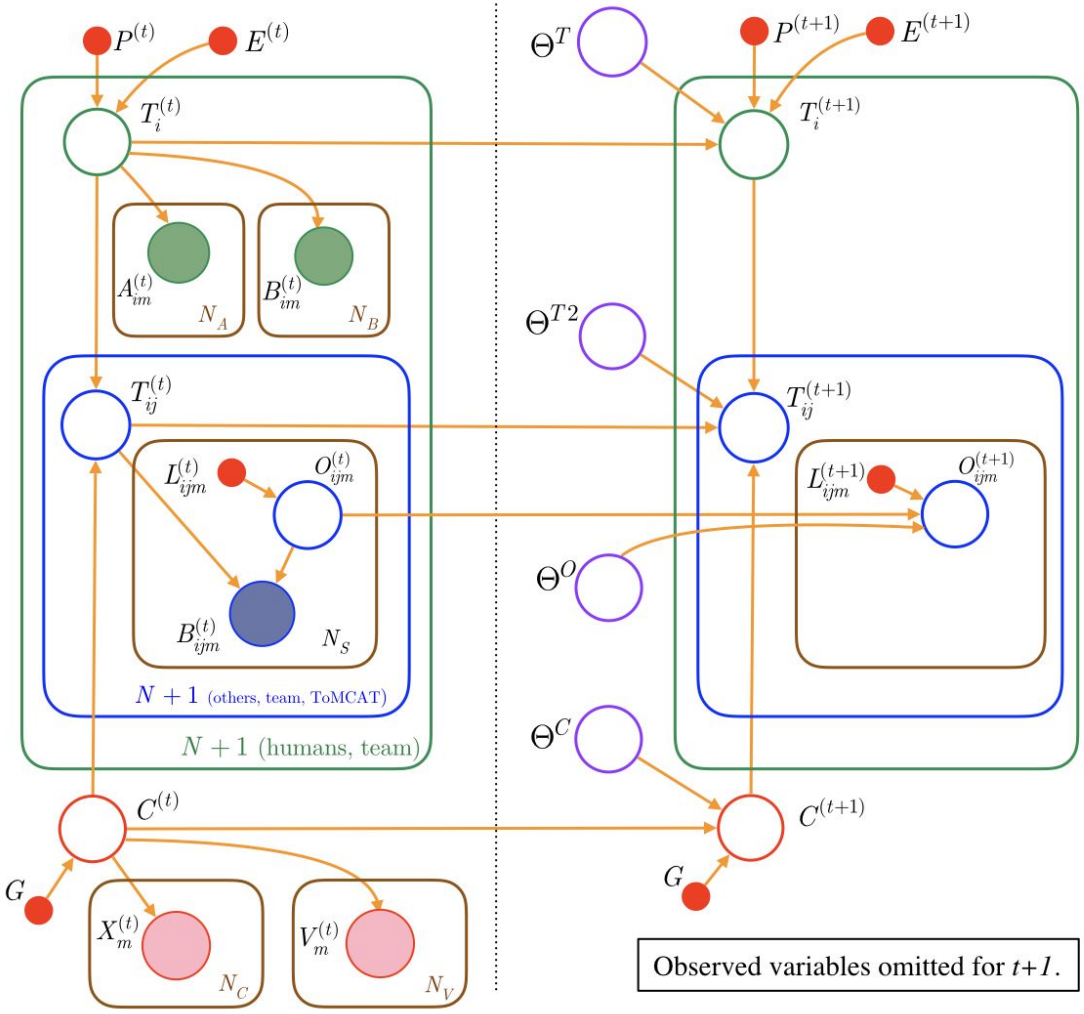
A_{im} $\Leftrightarrow$ Action by i with measure m

B_{im} $\Leftrightarrow$ Affective behaviour of i

B_{ijm} $\Leftrightarrow$ Affective behaviour of i towards j

X_m $\Leftrightarrow$ Coordination measures

V_m $\Leftrightarrow$ Team performance (value) measures



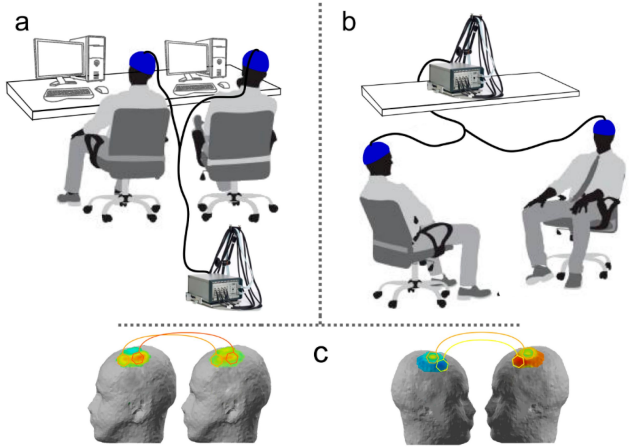
Clarifying possible misunderstandings related to using a DBN for ToMCAT

- We **can** take advantage of deep learning successes
  - For *higher-level* parts of the system deep learning **does not make sense**, and there will **not be enough data**.
  - For data representation in *lower-level* components, we expect that deep learning will be helpful
    - e.g., a CNN representation of an image that links to a more explanatory variable
    - Joint learning of a few network layers with the DBN likely makes sense here
- Mitigating concerns about slow inference
  - It is reasonable to expect our computers to work hard to get good answers to complex problems 😊
  - Just in-time-strategy
    - Inference processes can take advantage of hypotheses proposed by faster, less accurate, methods
    - Inference can improve such “data-driven” hypotheses, and the probabilistic model provides principled evaluation to select between them
    - This strategy works well with preset computation budget
  - Evaluation of multiple algorithms running on CPUs and GPUs
  - Good programming (we have lots of experience with large scale inference on PGMs)
  - Use capable hardware

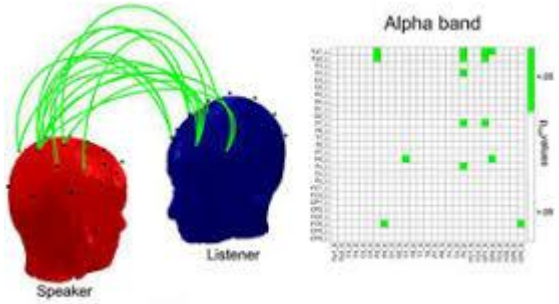
Why use brain hyper-scanning ?

- Team theory of mind and shared mental models live in a distributed space across the brains of the team members.
- So, if we want to study them, we should look at the team members' brains!
- Brain hyper-scanning refers to assessing multiple brains at once.
- Technological advances in the last 10-15 years have made that possible in ecologically valid settings.

fNIRS

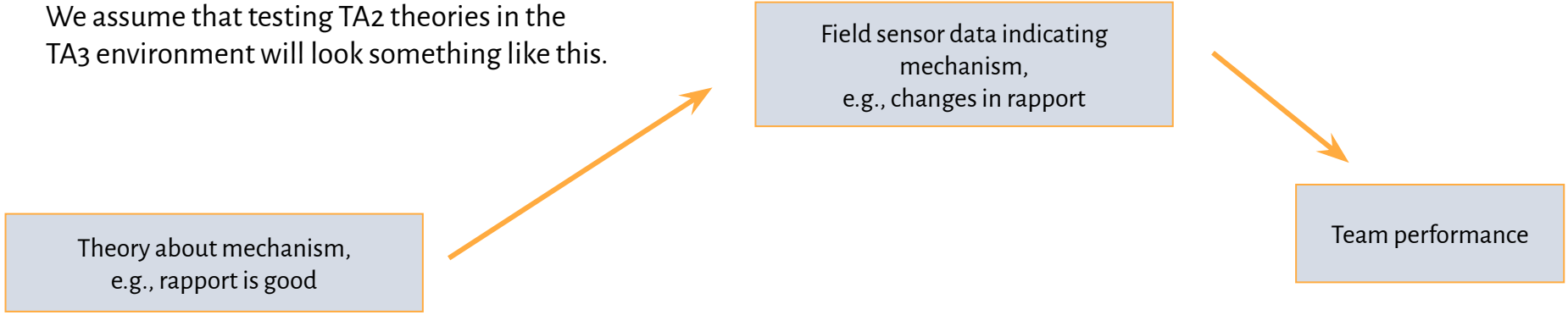


EEG



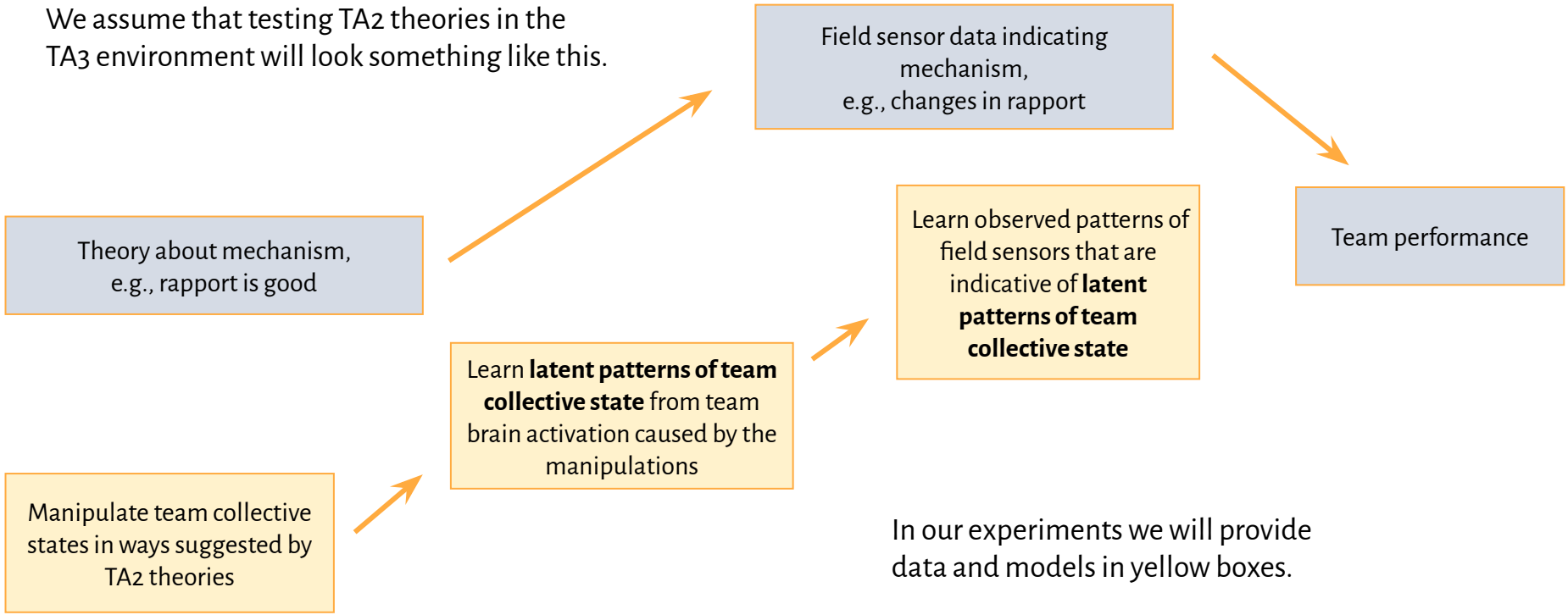
How will we support TA2 teams with brain hyper-scanning ?

We assume that testing TA2 theories in the TA3 environment will look something like this.



How will we support TA2 teams with brain hyper-scanning ?

We assume that testing TA2 theories in the TA3 environment will look something like this.



In our experiments we will provide data and models in yellow boxes.

Linking hyper-scanning together with other (cheaper) sensors and data sources

- The top of the "grounding" food chain for evaluation is performance and specific self-report
- Team theory constructs often will map to brain hyper-scanning
- We will map brain hyper-scanning to multimodal data from cheaper, more ambulatory measures
  - Key when hyper-scanning is not available or practical
- Plausible measures of success
  - (for TA2) Theory → brain → ambulatory → performance **is better** than other theories
  - (for us) Using brain space as a predictor **is comparable** to expensive expert human coding
    - IE, how does brain space capture "teaming" compared with human observers
  - (for both) Ambulatory → brain → theory → performance is better than ambulatory (alone), which could arguably win in the case of infinite data.

Measurement modalities

- A simultaneous combination of 4 person:
  - fNIRS
  - EEG
  - autonomic physiology (interbeat interval, skin conductance)
  - external surveillance (cameras on face, voice, eye tracking)
  - self report (during game playing, evaluation of one's own affective state while watching replay, surveys)

EEG and autonomic physiology provided by Brain Vision LLC and fNIRS provided by NIRx Medical Technologies. These companies both have outstanding reputations and work together routinely to provide integrated EEG/fNIRS hyper-scanning systems.

Eye tracking will a relatively inexpensive open source based product from pupil labs.

LANG LAB FOR FAMILY AND CHILD OBSERVATIONAL RESEARCH

Where will we do it?

The Janet and Barry Lang Laboratory for Family and Child Observational Research is a state-of-the-art lab designed for the collection of audio, visual, physiological data (EKG, impedance cardiography, respiration, blood pressure, skin conductance, and pulse amplitude) in studies of child, couple, group, and family interaction.

The Lang Laboratory offers:

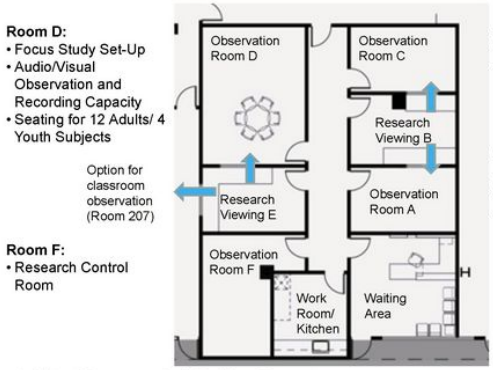
- Three rooms for observation
- One classroom for observation
- Two control rooms with video and audio equipment
- Variation in room size and set-up
- One-way mirrors
- Video and sound capabilities
- Tobii Studio Eye Tracke

Software available for use in the Lang Lab:

- The Observer XT 11.5
- Tobii Studio
- The Observer XT 10.5
- MPEG Recorder 2.0
- ClearView 2.7.1
- Sorenson Squeeze 4
- E-Prime 2.0
- Mindware BioLab 3.0

For more information on how to reserve a room in the Janet and Barry Lang Laboratory for Family and Child Observational Research, please click **here**.

The Lang Children and Family Observation Laboratory
McClelland Park – 650 North Park Avenue – Suite 215



Additional Resources Available Upon Request:
•Flip Charts
•Telephones With Conference Capability

Couple / Small Group Observational Room (Room C)

Large Group Observational Room / Zoom Videoconferencing Room (Room D)

What is fNIRS hyper-scanning ?

- Functional near-infrared spectroscopy.
- Non-invasive and relatively inexpensive way to measure brain activity.

- Based on the “hemodynamic response” (same as fMRI):
  - Neuronal activity is fueled by glucose metabolism in the presence of oxygen.
  - Increases in neuronal activity result in flooding of neuronal tissues with oxygenated hemoglobin (oxy-HB).
  - Bouts of activity result in a temporary increase in concentration of oxy-HB and a decrease in concentration of deoxygenated hemoglobin (deoxy-HB).

  - oxy-HB and deoxy-HB absorb and scatter near-infrared (NIR) light at different wavelengths in the range of 700-1000 nm.
  - Light emitters are placed on the scalp and radiate NIR into the head.
  - Differential absorption and backscattering of oxy-HB and deoxy-HB produce interpretable patterns of NIR light returning to the scalp, where it is measured with photodetectors.

- Scanning depth ~ 3 cm.
- Spatial resolution ~ 5 mm.
- Temporal resolution: Sampling rate = 100 hz. Actual image frame rate depends on application. Most common is ~ 3 -25 hz.

What is EEG hyper-scanning ?

- Electroencephalography.
- Also a non-invasive and relatively inexpensive way to measure brain activity.

- Assesses electrical activity in the brain.
  - Measured by electrodes on the scalp that record the electromagnetic field (in the direction perpendicular to the scalp) that is produced by a large population of neurons.
  - Two primary types of information: 1) different types of neural oscillations in the frequency domain (e.g., alpha, theta, etc. distinguished by different oscillatory frequencies), and 2) fluctuations connected to a time-locked event (e.g., event related potentials).

- Scanning depth and spatial resolution: Not well specified; generally not very precise and improving it is an open research area.

- Temporal resolution: There are a range of sampling rates. We would use ~ 7500 hz. Can assess electrical fluctuations in milliseconds.

Why use both EEG and fNIRS?

- EEG is an older technology, so it will provide a validity check for the newer fNIRS.
- They assess different aspects of brain function (electrical and hemodynamic) and it is unknown which will be a better indicator of team latent states. Quite likely, they will both provide relevant, but distinct and complementary, information.
- fNIRS provides spatial resolution that is lacking for EEG, providing more information about what team processes are involved (e.g., self-regulation, task-planning, motor-coordination, etc.).
- The faster time course of EEG may be useful for assessing coordinated team responses to external stimuli (e.g., stimuli that impacts each team member simultaneously and is not mediated through interpersonal behavior).
- The slower time course of fNIRS may be useful for assessing more complex team states that develop over seconds to minutes as information is passed between team members.

Existing literature on fNIRS and EEG hyper-scanning

- Paradigms are fairly limited:
  - Simple motor movements, e.g., button pushing, mimicking hand movements, with an emphasis on cooperating vs. competing.
  - Game theory contexts, again with an emphasis on cooperation vs. competition.
  - Music production.
- Results typically show synchronization increases in cooperative contexts and when trying to coordinate behavior (e.g., musicians coordinating based on a metronome) .
  - Often the results are specific to the prefrontal cortex, but earlier work only assessed that part of the brain, so this generality is questionable.
  - For EEG, results vary in terms of which frequency bands are involved.

Existing literature on fNIRS and EEG hyper-scanning

- Modeling approaches are very limited:
- Time-domain measures are typically used for fNIRS (e.g., cross-correlation, coherence, Granger correlation)
- Frequency domain measures are most used for EEG (e.g., partial directed coherence, frequency specific correlations).
- None of the methods used can distinguish different latent patterns of team states (e.g., in-phase, anti-phase, drifting, amplifying, damping, role differentiation or diffusion, etc.).
- Usually only for dyads, not teams, but there is a recent exception using network approaches to model the group as a whole.
- We could not find any work modeling fNIRS and EEG together from multiple people
    - There is a fair bit of work combining them in **individuals** for HCI generally showing they provide complementary information that increases prediction and classification.

Example to help internalize attributes of various measurements

The ToMCAT Team (so far!)

PIs

Kobus Barnard
Emily Butler
Clayton Morrison
Adarsh Pyarelal
Rebecca Sharp
Mihai Surdeanu
Marco Antonio Valenzuela-Escarcega

Graduate Students

Savannah Boyd
Loren Champlin
Harry Go
Ashley Kuelz
Paulo Soares
Manujinda Wathugala

Undergraduates

Aditya Banerjee
Lize Chen
Jiangfeng Li