galois

Leveraging Manifold Learning and Relationship Equity Management for Symbiotic Explainable Artificial Intelligence

Sourya Dey, Adam Karvonen, Ethan Lew, Donya Quick, Panchapakesan Shyamshankar, Ted Hille, Matt LeBeau, *Eric Davis**

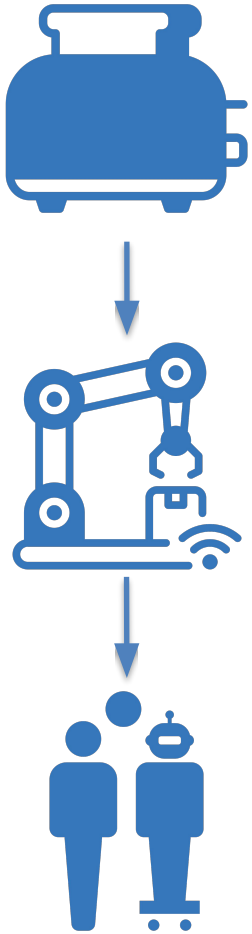
ewd@galois.com

Jul 2023

Interdependence in Human-Machine Teams

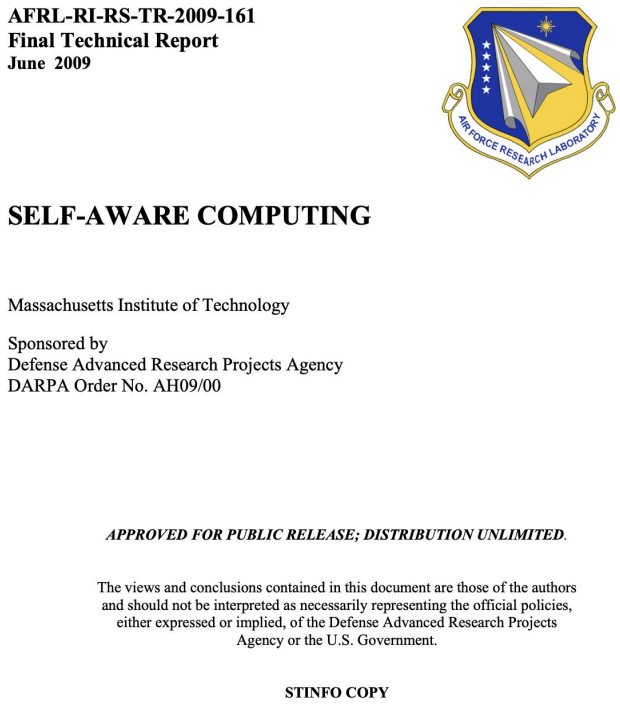
As the capabilities of automation have grown, there is a desire for them to function in ways beyond appliances, or automated performers, but as **interdependent teammates**.

AI need to do more than produce predictions from data. They need to understand their **interdependent purpose in joint activity**. They should **establish, maintain, and repair trust, loyalty**, and **seek to understand their shared task and the intent and abilities** of their human co-performers.



Self-Aware Computing

1. It is **introspective** or **self-aware** in that it can observe itself and optimize its behavior to meet its goals.
2. It is **adaptive** in that it observes the application behavior and adapts itself to optimize appropriate application metrics such as performance, power, or fault tolerance.
3. It is **self healing** in that it constantly monitors its resources for faults and takes corrective action as needed. Self healing can be viewed as an extremely important instance of self awareness and adaptivity.
4. It is **goal oriented** in that it attempts to meet a user's or application's goals while optimizing constraints of interest.
5. It is **approximate** in that it uses the least amount of precision to accomplish a given task. A self-aware computer can choose automatically between a range of representations to optimize execution -- from analog, to single bits to 64-bit words, to floating point, to multi-level logic.



AFRL-RI-RS-TR-2009-161
Final Technical Report
June 2009

SELF-AWARE COMPUTING

Massachusetts Institute of Technology

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. AH09/00

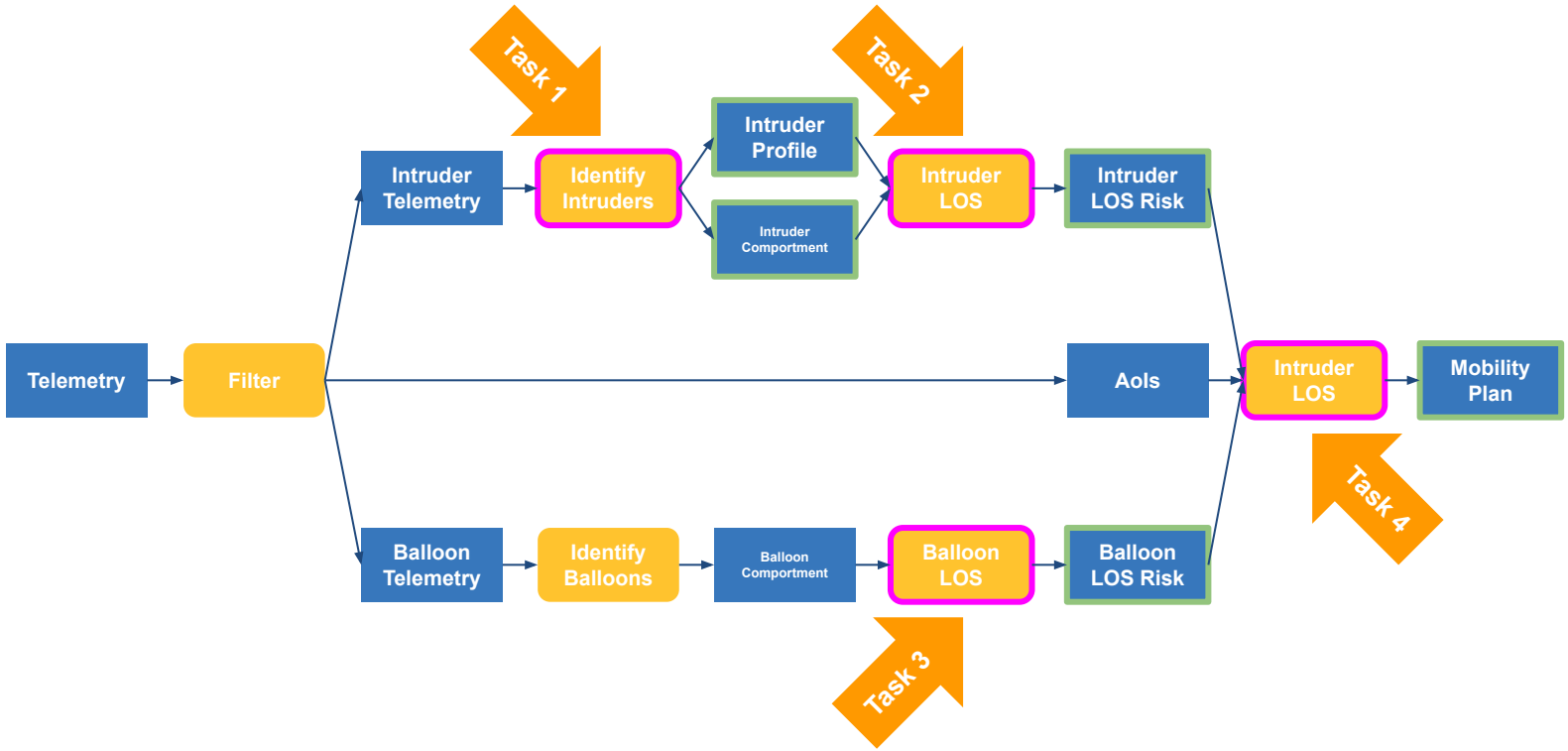
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

STINFO COPY

“Sense of Self” and “Theory of Mind”

Captures a prototype sense of self and theory of mind in its use of Joint Activity Graphs, and labeled functions.



Flexibility of JAG-based Reasoning

Use of Semantically Labeled Interchangable Reasoning

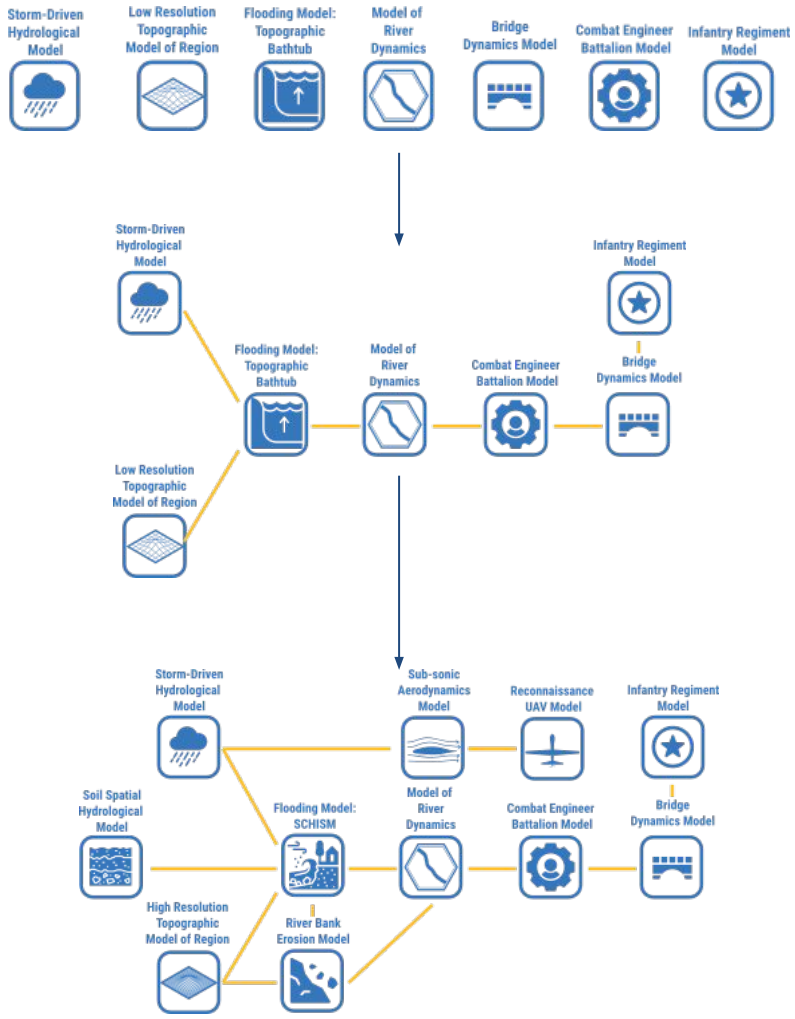
- Units of reasoning can come from varied producers.
- Varying quality and assumptions.
- Including surrogate models built with DeepKoopman

Adaptability

- Units of reasoning evolve over time
- Co-training can improve the results

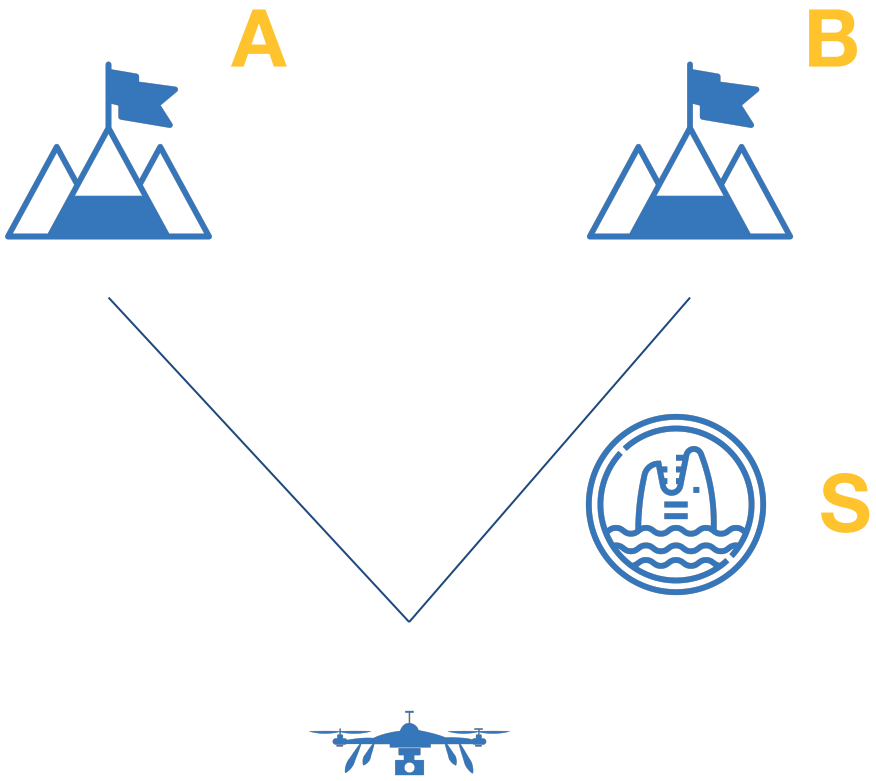
Model Resolution and "Cost"

- Can balance the temporal, execution, and co-performance risk using alternate pathways of reasoning and computation.



Modeling Commander's Intent

NATO's frameworks for Commander's Intent note that the key aspects that lead to success in joint operations are self-synchronization and understanding complex causes and effects.

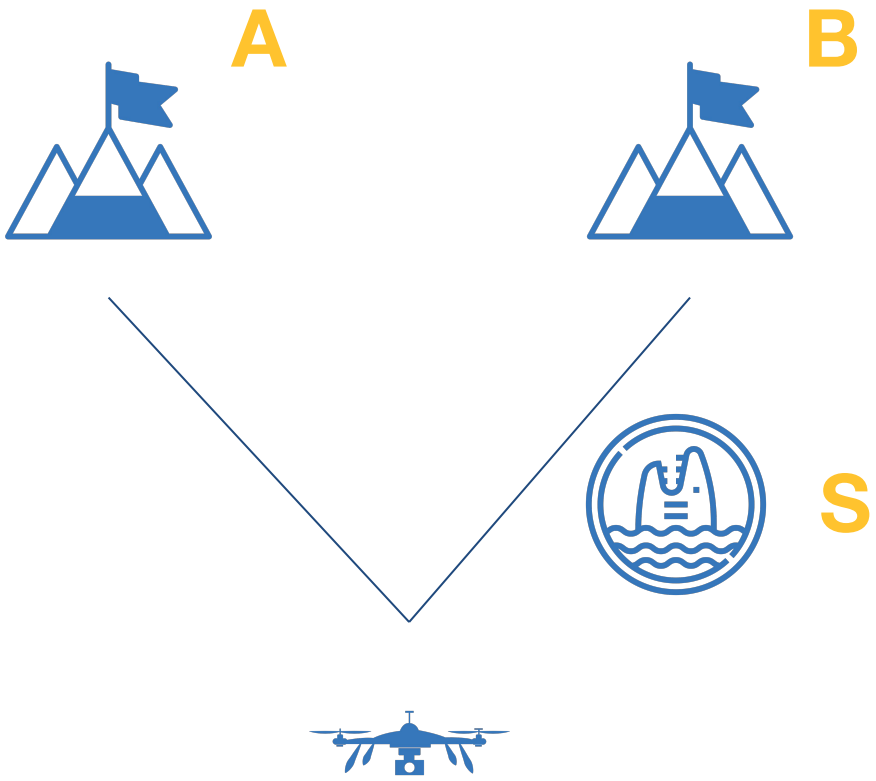


Modeling Commander's Intent

Which objective should we attempt to achieve?

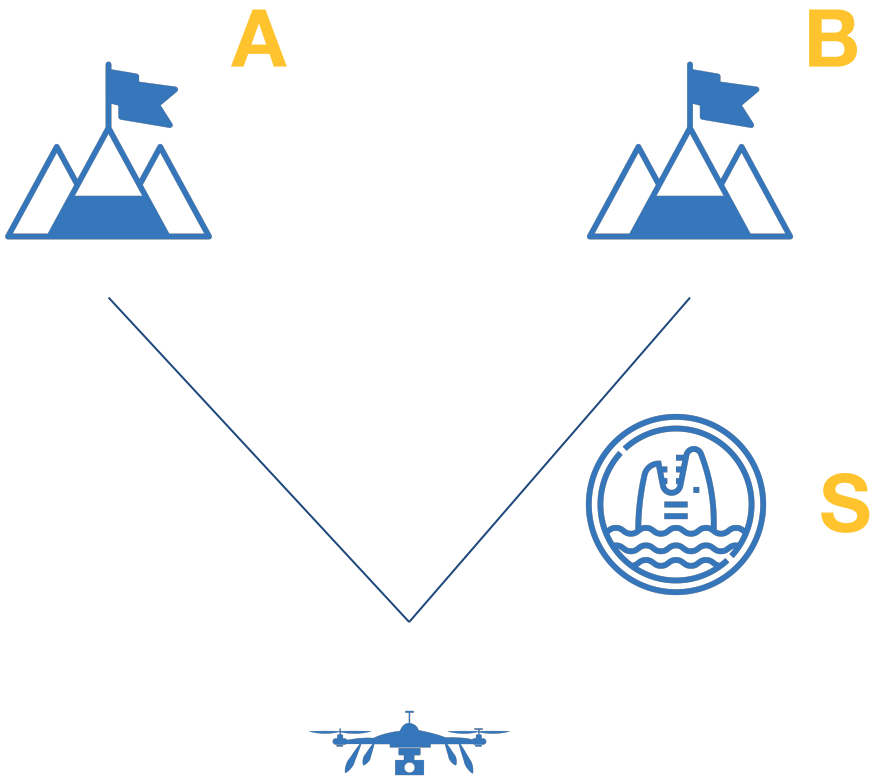
Objective A is achievable without complications.

Objective B forces us to face hazard S.



Modeling Commander's Intent

Self-synchronization *(A. Alberts, Gartska, & Stein, 1999)* is a term used to describe the operational patterns displayed by entities in the absence of traditional hierarchical mechanisms for command and control.

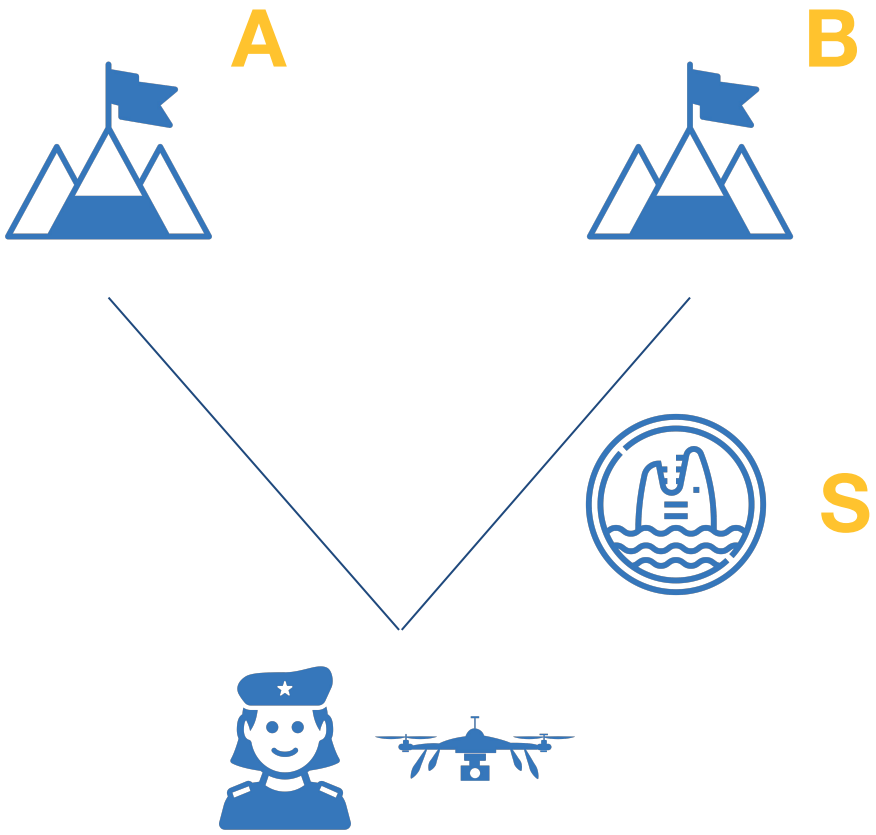


Modeling Commander's Intent

When a commander is present, we can clarify objectives during uncertainty.

Uncertainty will almost always occur naturally.

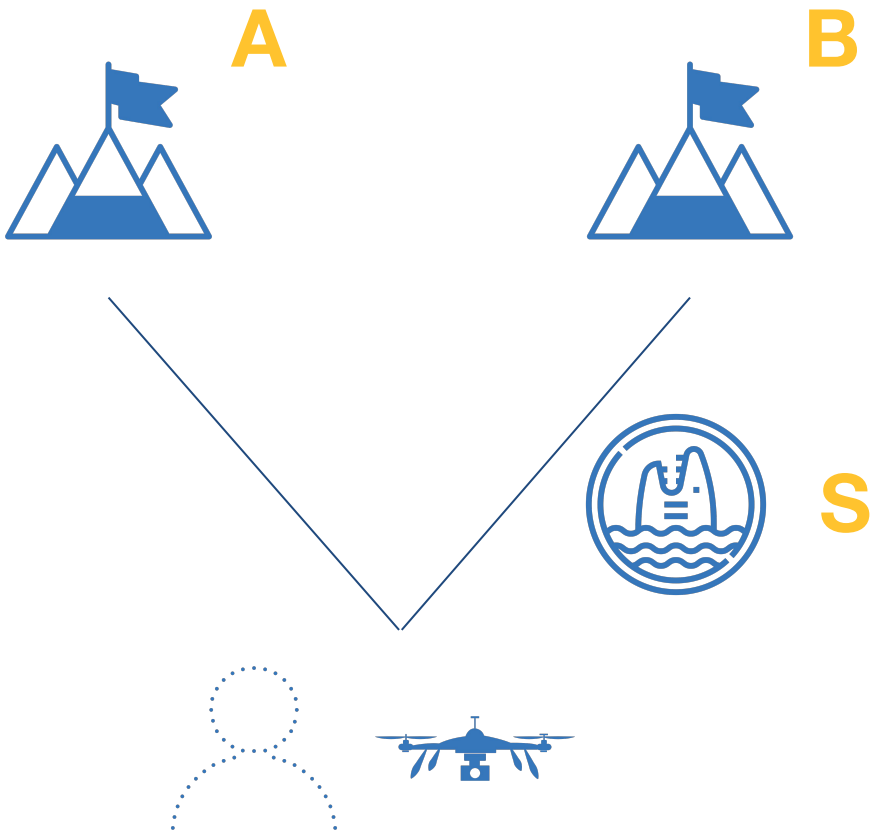
Assuring that these decisions are reflective of the goals, missions, and objectives of their human co-performers and real commanders is ***a notoriously difficult problem that cannot be solved with explicit constraints and objectives alone*** *(Russell, 2019; Tegmark, 2017).*



Modeling Commander's Intent

When the commander is absent, we must rely on self-synchronization.

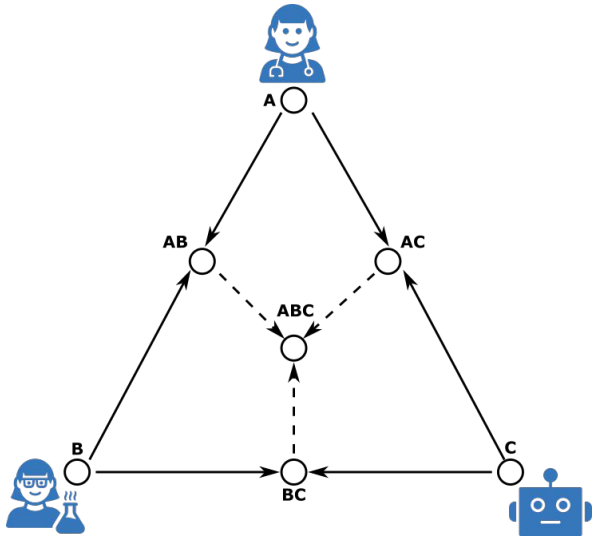
Modern AI does not have this capability.

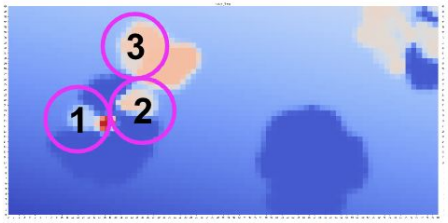


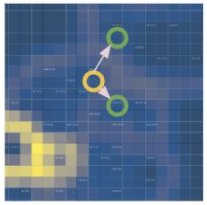
Modeling Commander's Intent

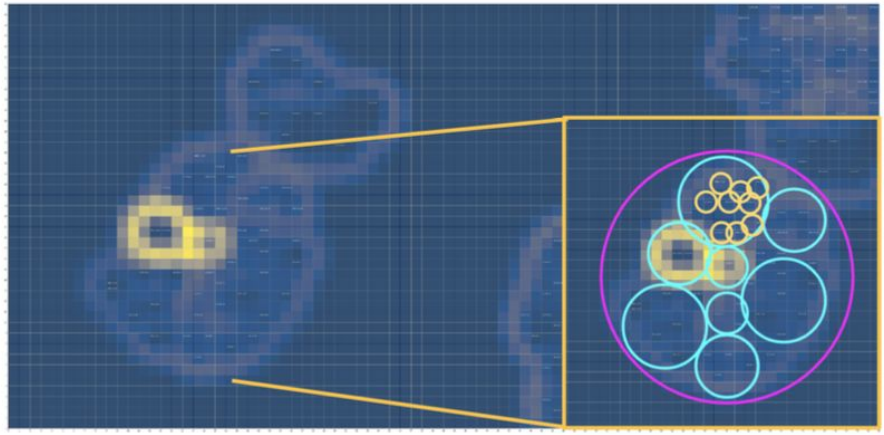
We can now estimate the underlying topology.

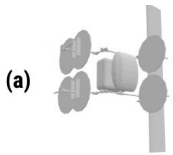
Initial experiments with three synthetic sets of commander's intent on sub-diagnosis for heart disease showed alignment with ~98% accuracy.











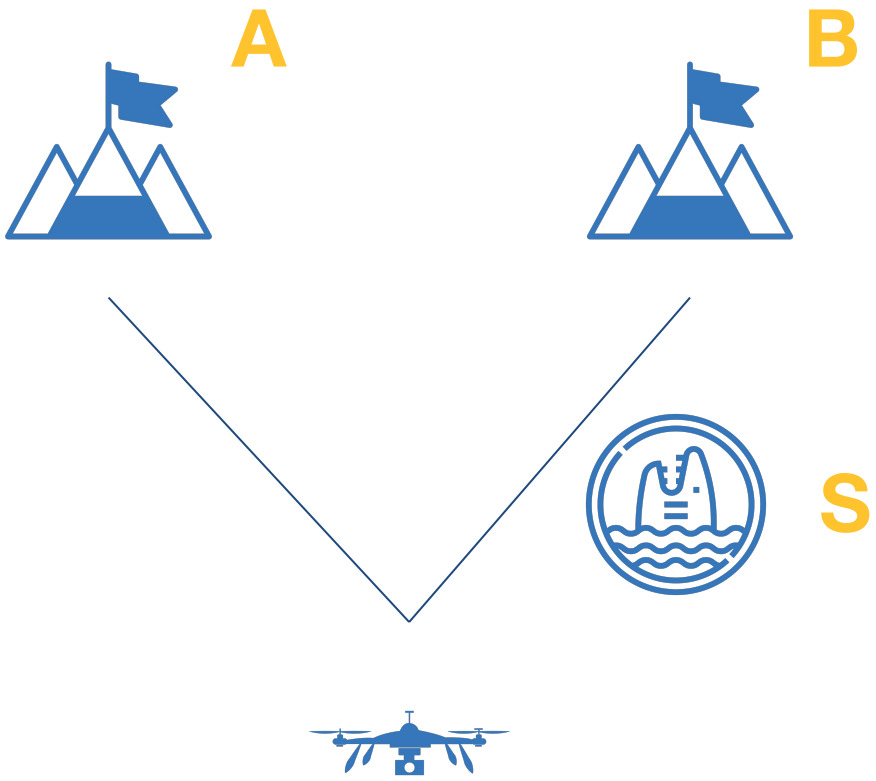
Modeling Commander's Intent

Conceptual pacts tell us what a Commander **thinks**. The estimation of the **topology** allow us to compare items, and place them in the topology to generate Euclidean distance metrics.

What does a "successful mission" look like?

What does "acceptable losses" mean in a situation?

We can model this synthetic commander in the field, during self-synchronization.



| galois |

Thank you!!!

Dr. Eric Davis, Principal Scientist

ewd@galois.com
ewd@symbioticai.org

Galois [gal-wah] Named after French mathematician Évariste Galois

www.galois.com