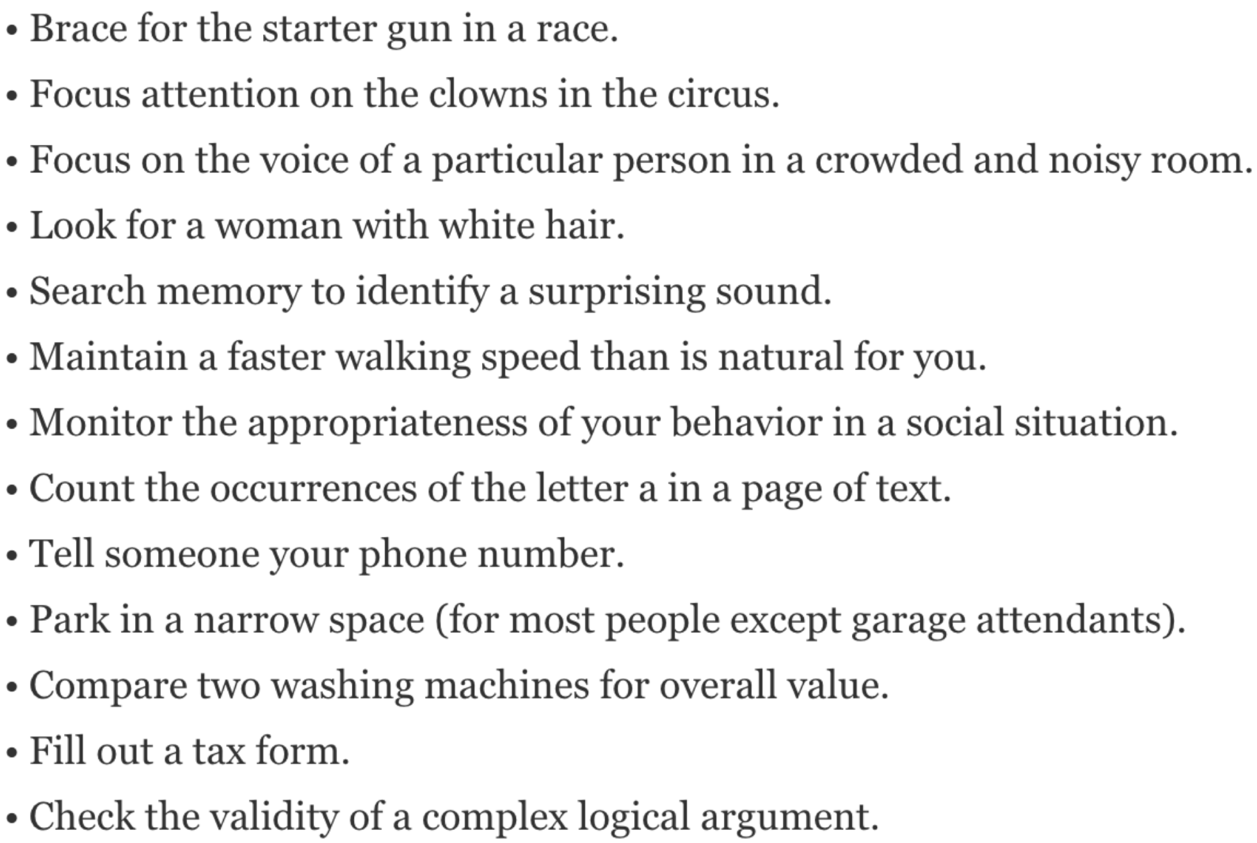
Research Topics in S1 and S2

Sargur N. Srihari
srihari@buffalo.edu

System1 Tasks

- Detect that one object is more distant than another.
- Orient to the source of a sudden sound.
- Complete the phrase “bread and . . .”
- Make a “disgust face” when shown a horrible picture.
- Detect hostility in a voice.
- Answer to 2 + 2 = ?
- Read words on large billboards.
- Drive a car on an empty road.
- Find a strong move in chess (if you are a chess master).
- Understand simple sentences.
- Recognize that a “meek and tidy soul with a passion for detail” resembles an occupational stereotype.

System 2 Tasks



- Brace for the starter gun in a race.
- Focus attention on the clowns in the circus.
- Focus on the voice of a particular person in a crowded and noisy room.
- Look for a woman with white hair.
- Search memory to identify a surprising sound.
- Maintain a faster walking speed than is natural for you.
- Monitor the appropriateness of your behavior in a social situation.
- Count the occurrences of the letter a in a page of text.
- Tell someone your phone number.
- Park in a narrow space (for most people except garage attendants).
- Compare two washing machines for overall value.
- Fill out a tax form.
- Check the validity of a complex logical argument.

Topics*

1. Capabilities of Systems 1, 2
2. Is System 2 inherently sequential?
3. Metrics to evaluate performance of Systems 1, 2
4. Introspection in Systems 1, 2
5. Switching between Systems 1,2
6. How to reason when using both Systems 1.2
7. Ethics theories tied to Systems 1,2
8. Definition of abstraction: attention, knowledge graphs
9. Combining multi-agent systems
10. Architectural choices for supporting this vision of AI

- * arXiv:2010.06002v2 [cs.AI] 15 Dec 2020

Proposal: Advances in AI

- Research direction to advance AI
  - Cognitive theories of human decision making in AI
- Gain insights into capabilities that are still lacking AI
  - adaptability
  - generalizability
  - common sense
  - causal reasoning
- To embed these causal components into AI systems

Overall Vision

- Many AI advances in capabilities, but pertain to narrow AI:
  - Image interpretation
  - Natural language processing
  - Label classification
  - Prediction
- They are tied to
  - Improved algorithms
  - Huge data sets
  - Computational power
- If we compare to human capabilities
  - Generalizability
  - Robustness
  - Explainability
  - Causal analysis
  - Abstraction
  - Common sense reasoning
  - Ethics reasoning
  - Complex and seamless integration of learning and reasoning supported by both implicit and explicit knowledge

Whether end-to-end neural networks suffice?

- Whether we need to integrate machine learning with symbolic and logic based techniques
- We believe that the integration route is the more promising
- Due to several reasons

Supporting statements/work

- Bengio 2017
  - Necessary to generalize from raw data

    to a ”consciousness stream” of few a concepts related to each other
- Marcus 2020
  - Explicit knowledge, symbols, and reasoning

    should be used to improve the robustness of current AI systems
- d’Avila Garcez and Besold 2019
  - Building hybrid systems that use both ML and symbolic reasoning techniques, employing neuro-symbolic AI approach

Arguments in favor

- Humans have evolved to obtain these competencies
- We propose to use human reasoning to raise fundamental issues that are lacking
- Special focus on thinking fast and thinking slow

What needs to be done

- It is possible to draw a parallel between main lines of work in AI:
  - Machine learning
  - Symbolic logic reasoning
- Between data-driven and knowledge driven AI
- Perception-based, such as seeing and reading, lack in basic notions of causality
- Basic notions of AI techniques based on
  - logic, search, optimization, planning
- And employing
  - explicit knowledge
  - symbols, and
  - high-level concepts

Research Questions1

Capabilities of Systems 1, 2

- Identifying capabilities of Systems 1,2
- What features would they use?
- While System 2 is sequential it does not necessarily have to be this way in a machine

Research Questions 2.

Is System 2 inherently sequential?

- System 2 is sequential, does a machine system 2 also have to be sequential?
- Should we exploit parallel threads performing System 2 reasoning?
- Would this, together with the greater computing power of machines compared to humans, compensate for the lack of other capabilities in AI?

Research Questions 3
Metrics to evaluate performance of Systems 1, 2

- System 1: Accuracy, Precision/Recall/F-measure
- System 2: Correctness of conclusion
- In a combined system switching between the two
  - CHECKLIST is a method to evaluate a Sentiment Analysis system

Capability	**Min Func Test**	**INVariance**	**DIRectional**
Vocabulary	Fail. rate=15.0%	16.2%	34.6%
NER	0.0%	20.8%	N/A
Negation	76.4%	N/A	N/A
….			

4. Introspection in Systems 1, 2

- How do we define AI's introspection in terms of I-consciousness and M-consciousness
  - Can M-consciousness alone solve complex problems that in humans need also sophisticated levels of I (Information)-consciousness?
  - How is introspection linked to autonomy and human-machine teaming?
  - Is introspection a binary concept or should it have several levels?
    - If in levels, how can we associate different capabilities to the various levels?
  - Should we have models of the agents' minds (those supporting M-consciousness) at different levels of precision/fidelity?
- We may need to make a high-stakes decision
  - A mistake may confer dangerous consequence
  - Or explainable to a third party

5. Governance: Switching between Systems

- AI system should be able to recognize when it needs to deploy capabilities from system 1 or system 2, and also when to switch between the two systems
  - In humans, such a switch depends on many factors, including framing, priming, and the context of the decision environment.
  - In an AI system, there seems to be other factors that make up the sufficient conditions for this switch, e.g., when system 1 cannot find a solution
- We may need to switch when we are facing a very high stake decision:
  - A mistake may confer dangerous consequences according to some high-weight criterion.
  - Or also when the decision needs to be explainable to a third party.

6. Reasoning when using both Systems 1.2

- When there are two or more diverging policies or heuristics offered up by system 1, one needs to carefully explore the main pros and cons, as well as the possible consequences of our actions, before deciding on the policy to adopt, or we need to define a new policy
- Careful reasoning, based on inference and counter-factuals, seems to be a system 2 capability, supported by system 1's heuristics to reduce the search space

7. Ethics theories tied to Systems 1,2

- Deciding if a certain action is morally acceptable or not
- How humans switch between these different approaches in
  - Judging the morality of an action (Rossi and Loreggia 2019; Awad et al. 2020)
  - How divergence can be defined in an ethical context (Loreggia et al. 2018a)
  - How to define methodologies for ethical reasoning in AI (Loreggia et al. 2018b,c, 2019)

8. Abstraction, generalization and knowledge

- In order to adapt to a new environment, an agent need to be aware of its competencies
- How to deploy its skills and problem solving capabilities, were possibly acquired in another environment, into the new one
  - Recognize similarities and differences between the two environments, and to use this knowledge to decide what to temporarily forget during the abstraction step
- How do we define abstraction / generalization mechanisms that are guided by a notion of attention and pass from the raw data level to a more abstract level?

Abstraction: attention, knowledge graphs

- This conjecture is aligned with the so-called consciousness prior theory (Bengio 2017).
- Existing and well established abstraction theories (Cousot and Cousot 1977) and studies of various forms of abstraction in AI (Zucker2003) can be useful, coupled with attention mechanisms(Vaswani et al. 2017)
- What does it mean for knowledge to be explicit:
  - Is it related to the presence of metadata, structured knowledge graphs, or language-related entities?

9. Combining multi-agent systems

- Multi-agent epistemic reasoning is what humans engage on when making complex decisions
- In a multi-agent view of several AI systems communicating and learning from each other, how to exploit/adapt current results on epistemic reasoning and planning to build/learn models of the world and of others?

10. Architectures for supporting this vision of AI

- AI systems should include several independent simple components, to be triggered when needed according to a governance model
- Individual agents focus on specific skills and problems, act asynchronously, contribute to building models of the world, of other AI systems, and of self, and can be combined in many ways
- What architectural choices best support the above vision of the future of AI

SOFAI architecture

- Solve AI’s is to scale the models:

  Create bigger neural networks,

  Gather larger datasets,

  Use larger server clusters,

  Train the reinforcement learning algorithms for longer hours
- SOFAI uses *meta-cognition* to arbitrate between different modes of inference
  - Improve efficiency in using data and compute resources
- Uses Reinforcement Learning

Solving a real World problem

- A 2–d grid world problem
- Based on *Multi-alternative Decision Field Theory*

Multi-Alternative Decision Field Theory

- Personal Evaluation
  - Human iterative choice as cumulative process
  - Given a set of Options and Attributes
    - Tastes and Attributes

$$M = \begin{vmatrix} 1 & 5 \\ 5 & 1 \\ 2 & 3 \end{vmatrix}$$

Attention Weights

- Attention weights
- $\mathbf{W}(\mathrm{t})= [1,0]$ or $\mathbf{W}(\mathrm{t})= [0,1]$
- w_1= $0.55,$ w_2= 0.45

Contrast Matrix

Contrast an option wrt other options

Three options shown

$$C = \begin{vmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{vmatrix}$$

Valence Vector

At any moment of time,
each alternative in the choice set is associated with a valence value

- $\mathbf{V}(\mathrm{t})=\mathrm{C} \times \mathrm{M} \times \mathrm{W}(\mathrm{t})$
- $\mathrm{W}(\mathrm{t})=[0,1]$, $\mathrm{V}(\mathrm{t})=[1\text{-}7/2,\ 5\text{-}3/2,\ 2\text{-}6/2]$

Feedback Matrix

How the accumulated preferences affect the preferences computed at the next iteration

This lateral inhibition decreases as the dissimilarity between options increases

$$S = \begin{vmatrix} +0.9000 & 0.0000 & -0.0405 \\ 0.0000 & +0.9000 & -0.0047 \\ -0.0405 & -0.0047 & +0.9000 \end{vmatrix}$$

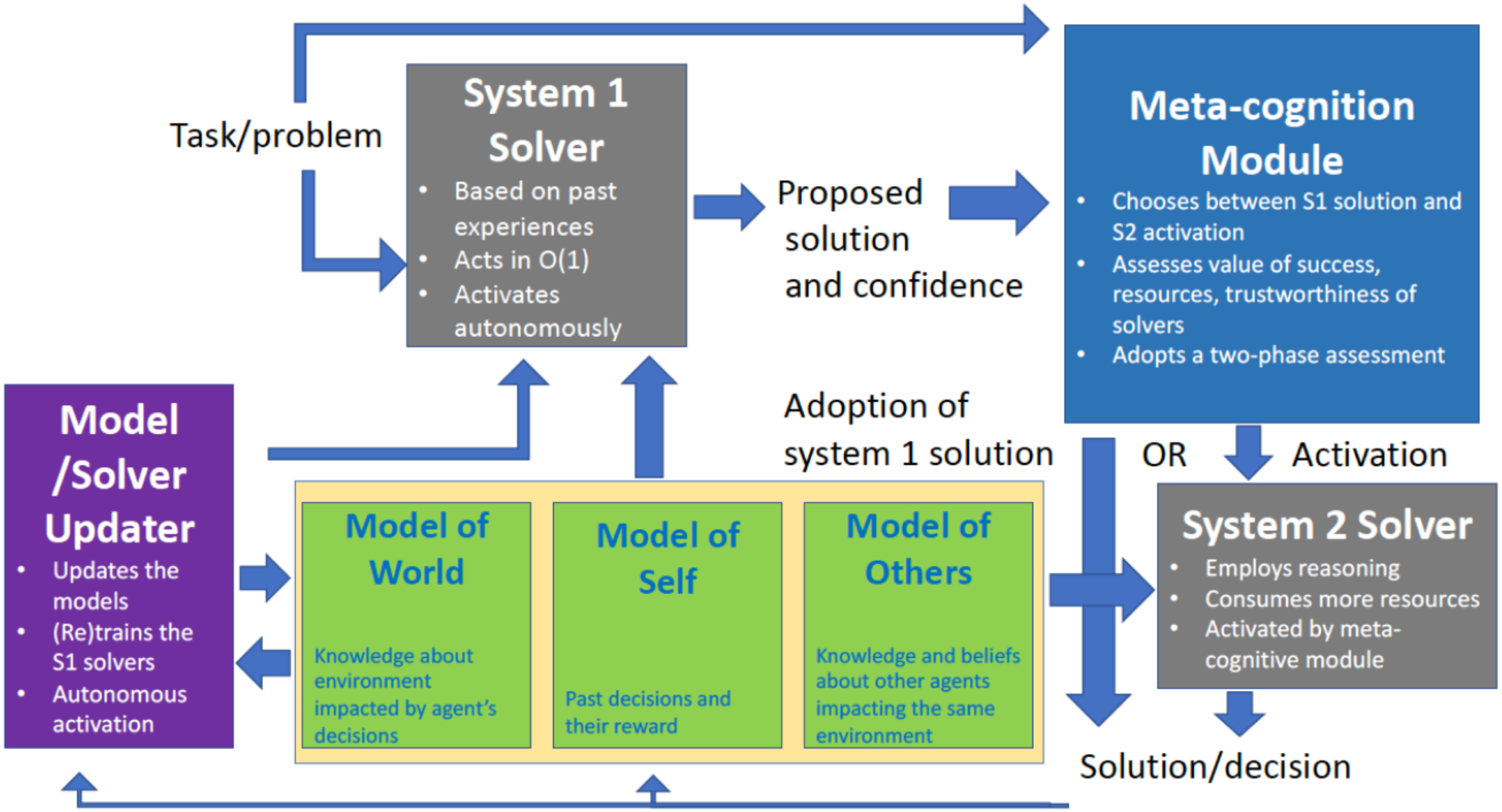
Preference for each alternative

- $\mathrm{P(t+1)} \; = \mathrm{S}$ x $\mathrm{P(t)} + \mathrm{V(t+1)}$
- Starting at $\mathrm{P(0)=0}$, preferences are accumulated

The SOFAI architecture

- Incoming problems are initially handled by System1 solvers
- Analogous to what is done by humans

System 1 and System 2



Task/problem

System 1 Solver

- Based on past experiences
- Acts in O(1)
- Activates autonomously

Proposed solution and confidence

Meta-cognition Module

- Chooses between S1 solution and S2 activation
- Assesses value of success, resources, trustworthiness of solvers
- Adopts a two-phase assessment

Adoption of system 1 solution

OR

Activation

System 2 Solver

- Employ reasoning
- Consumes more resources
- Activated by meta-cognitive module

Solution/decision

Model /Solver Updater

- Updates the models
- (Re)trains the S1 solvers
- Autonomous activation

Model of World

Knowledge about environment impacted by agent's decisions

Model of Self

Past decisions and their reward

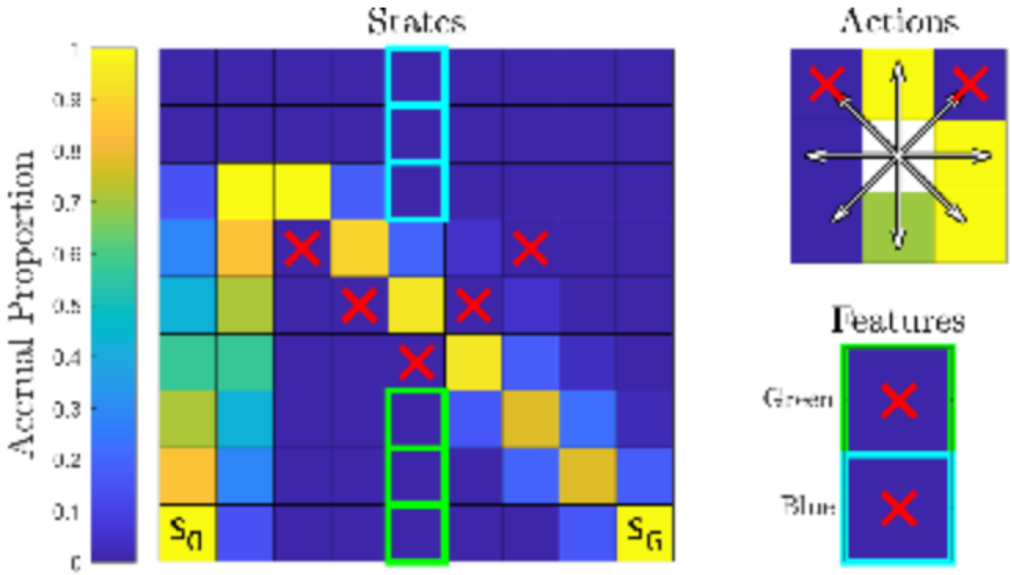
Model of Others

Knowledge and beliefs about other agents impacting the same environment

Choice between S1 and S2

- Once the S1 solver has solved the problem, the proposed confidence level is available to the Metacognitive model
- At this point the Metacognitive model starts to choose between the S1 solver or activating the S2 solver

Problem Domain



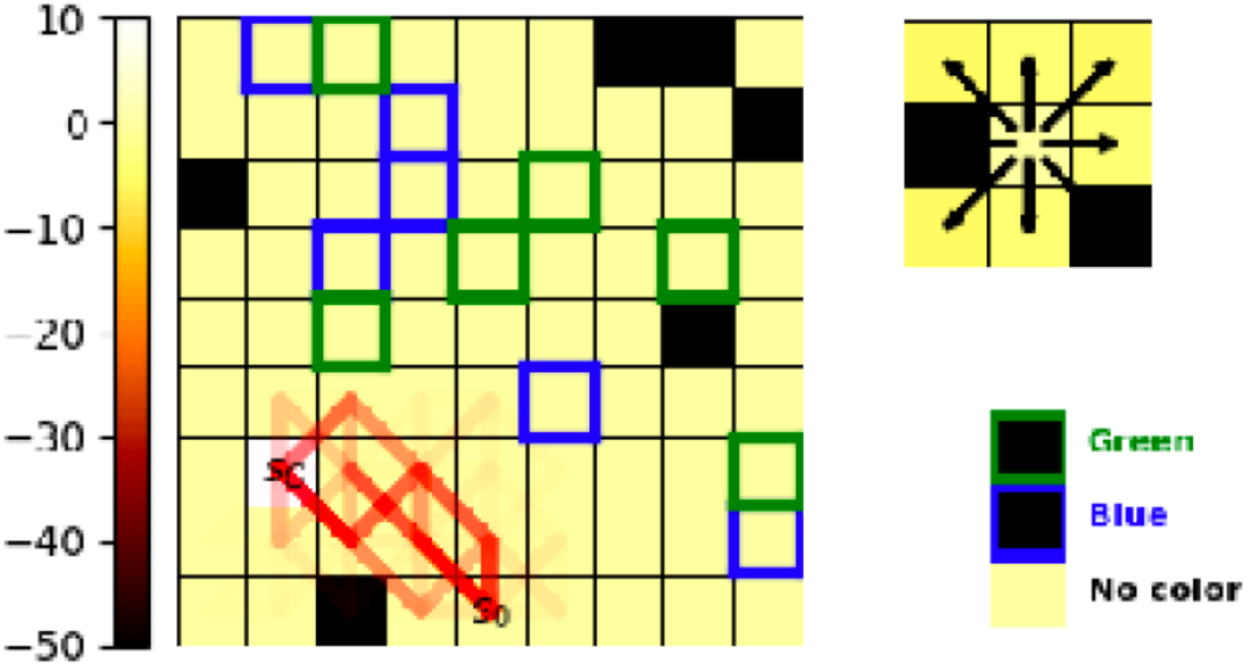
9 x 9 grid and the task is to generate a sequence from the start to the goal from one state to the next while minimizing the penalties incurred

S_o and S_G are start and goal

Different paths are high-lighted

X's and high-lighted cells represent obstacles

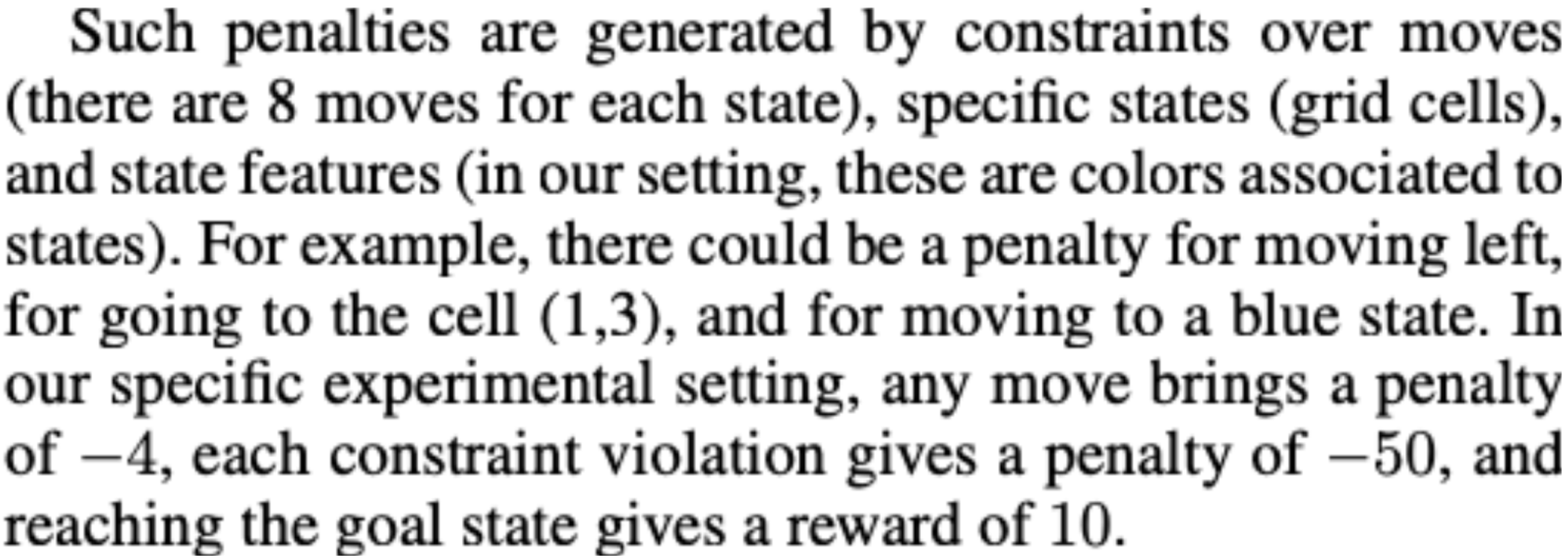
Problem Domain



Trade-off between S1 and S2

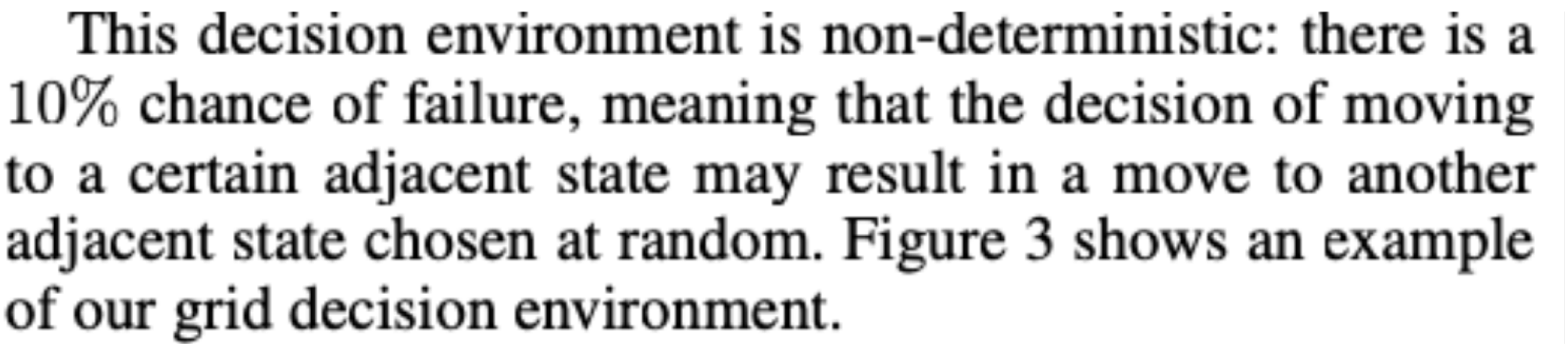
- S1 makes this available for short-term gain
- For long-term S2 should be available

Penalty for moves



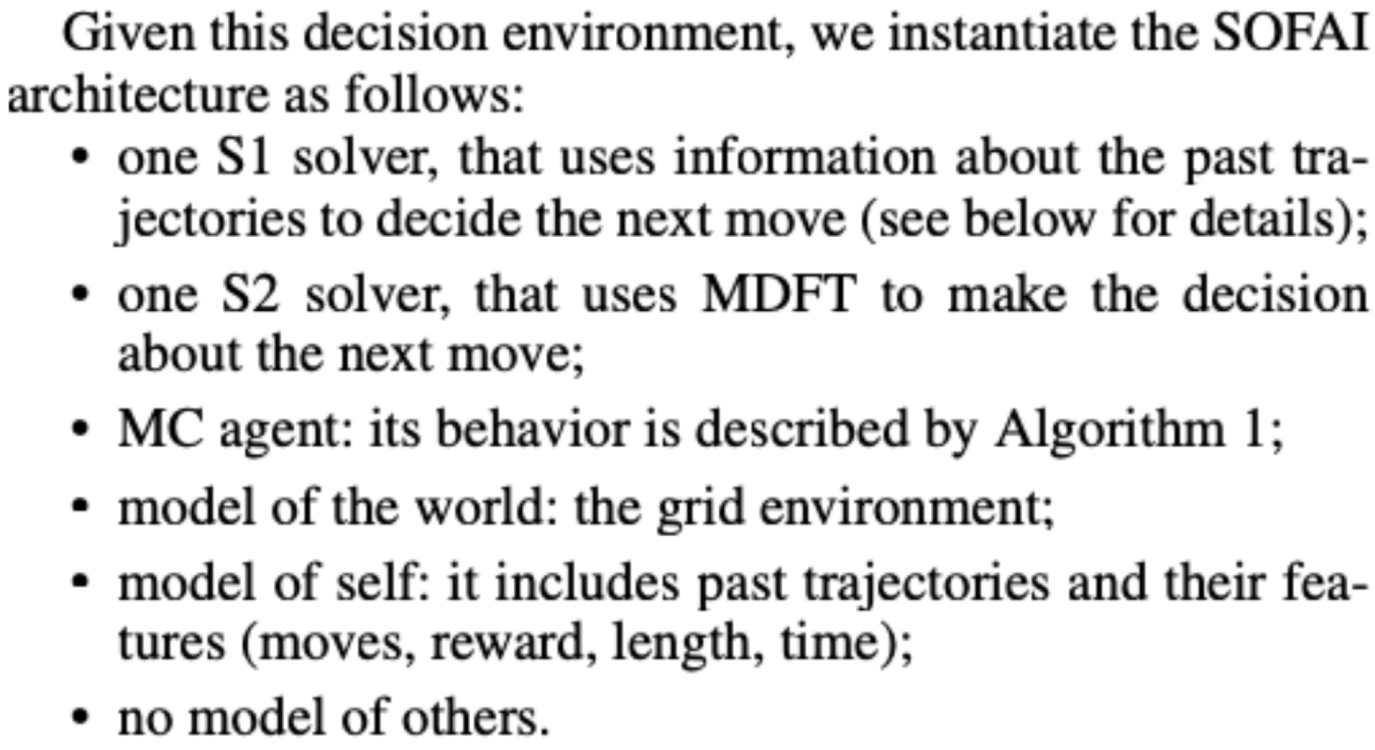
Such penalties are generated by constraints over moves (there are 8 moves for each state), specific states (grid cells), and state features (in our setting, these are colors associated to states). For example, there could be a penalty for moving left, for going to the cell (1,3), and for moving to a blue state. In our specific experimental setting, any move brings a penalty of -4, each constraint violation gives a penalty of -50, and reaching the goal state gives a reward of 10.

Non-deterministic decision environment



This decision environment is non-deterministic: there is a 10% chance of failure, meaning that the decision of moving to a certain adjacent state may result in a move to another adjacent state chosen at random. Figure 3 shows an example of our grid decision environment.

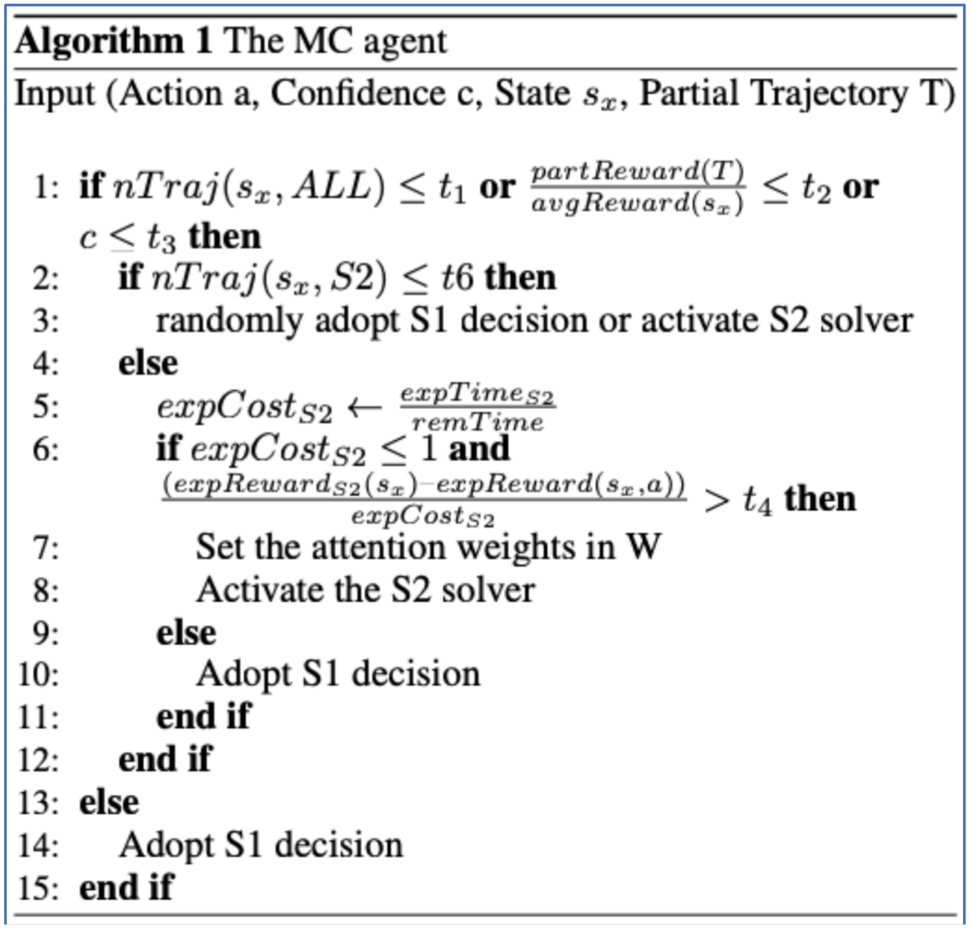
Instantiation of the SOFAI



Given this decision environment, we instantiate the SOFAI architecture as follows:

- one S1 solver, that uses information about the past trajectories to decide the next move (see below for details);
- one S2 solver, that uses MDFT to make the decision about the next move;
- MC agent: its behavior is described by Algorithm 1;
- model of the world: the grid environment;
- model of self: it includes past trajectories and their features (moves, reward, length, time);
- no model of others.

MetaCognitive Phase



Algorithm 1 The MC agent

Input (Action a, Confidence c, State s_x, Partial Trajectory T)

```
if $nTraj(s_x, ALL) \leq t_1$ or $\frac{partReward(T)}{avgReward(s_x)} \leq t_2$ or $c \leq t_3$ then
    if $nTraj(s_x, S2) \leq t6$ then
        randomly adopt S1 decision or activate S2 solver
    else
        $expCost_{S2} \leftarrow \frac{expTime_{S2}}{remTime}$
        if $expCost_{S2} \leq 1$ and $\frac{(expReward_{S2}(s_x) - expReward(s_x, a))}{expCost_{S2}} > t_4$ then
            Set the attention weights in W
            Activate the S2 solver
        else
            Adopt S1 decision
        end if
    end if
else
    Adopt S1 decision
end if
```

Compute a confidence as follows:

$c(s_x,\ a) = sigmoid(\ (r - 0.5)/(\sigma + 1e - 10))\ )$

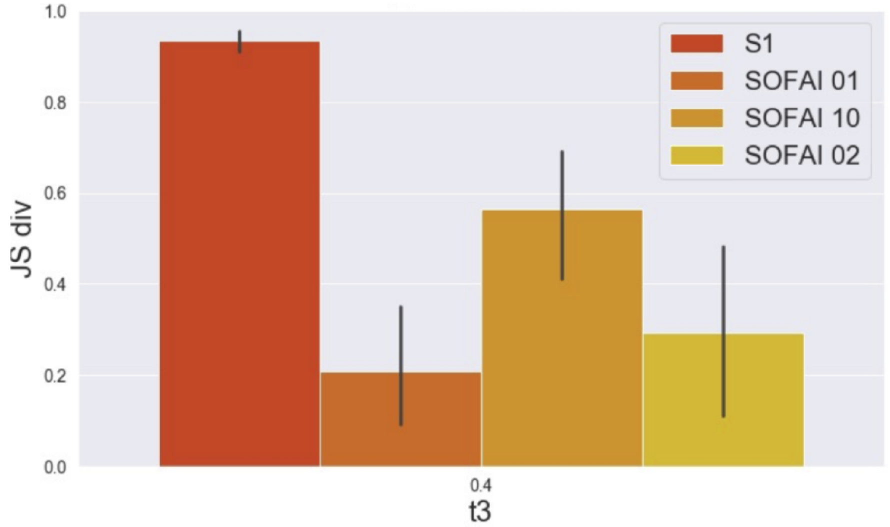
where σ is the standard deviation of the rewards in s_x taking an action a,

r is the probability of taking action a in state s_x

Experimental Results

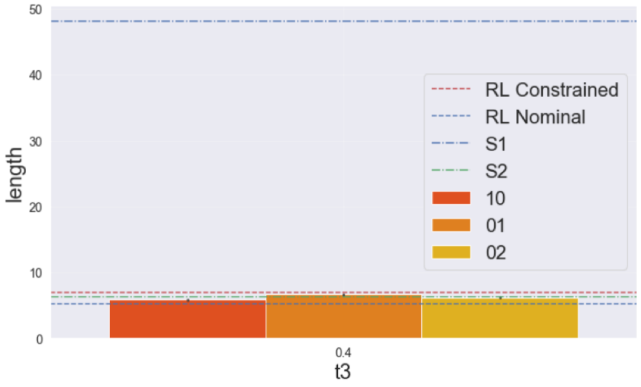
- Random 10 grids, and for each grid the initial and final states, 2 actions, 6 states, 12 state features (6 green and 6 blue).
- For each grid: two reinforcement learning agents
- S1 solver
- S2 solver
- Agent will be the provider of human-like trajectories
  - Each agent generates 1000 trajectories.
- We report the results for: $\mathrm{t}_1 = 200$, $\mathrm{t}_2 = 0.8$, $\mathrm{t}_3 = 0.4$, $\mathrm{t}_4 = 0$, $\mathrm{t}_6 = 1$.

JS Divergence (Between set of trajectories and the other systems)

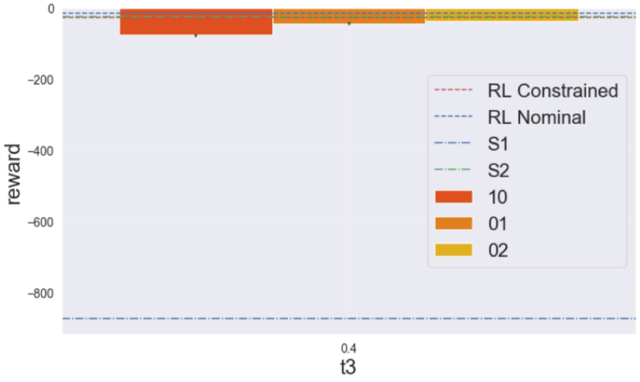


Divergence between components of S1 is higher than that between components of SOFAI. Implies less variation in S2.

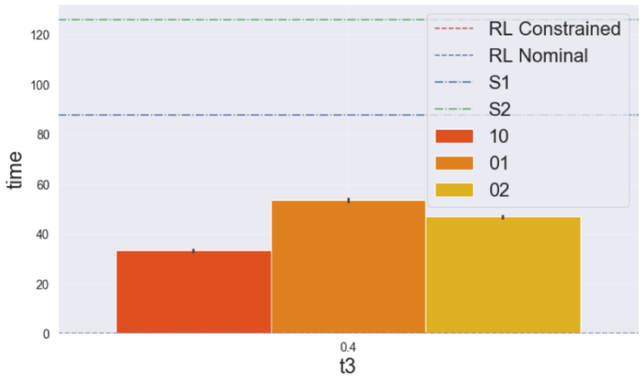
Comparison of three versions of SOFAI



Average Length



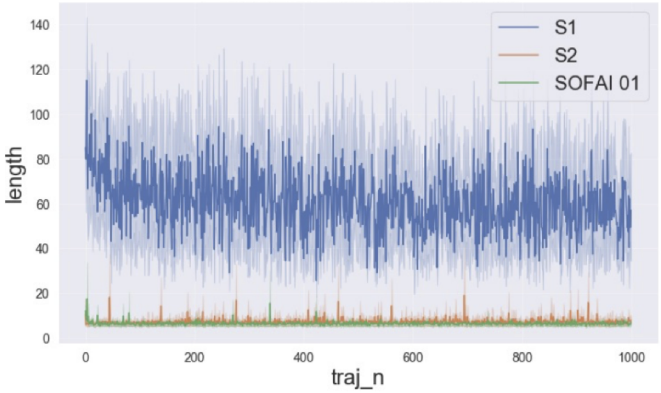
Reward



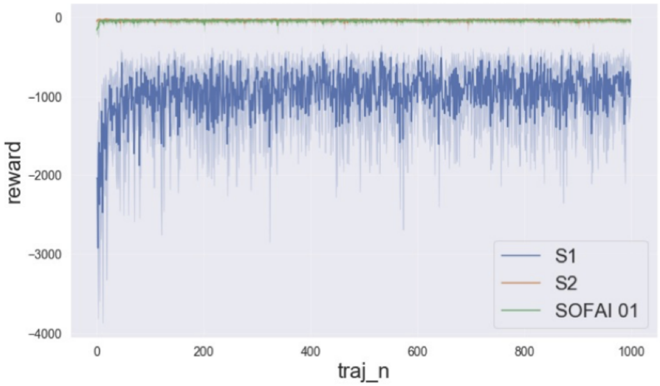
Time

(Aggregated over 10 grids and 1000 trajectories)

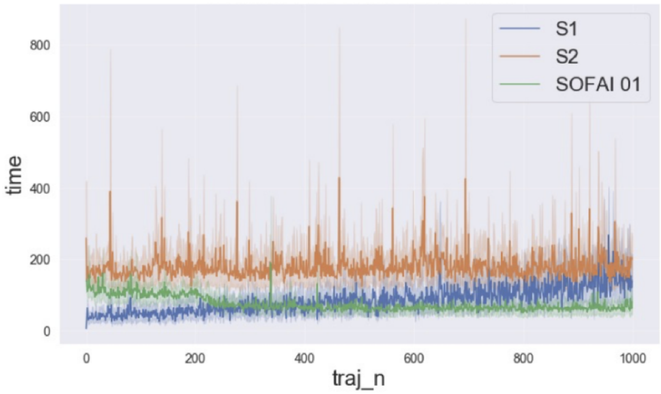
Length, Reward and Time comparing SOFAI to S1 and S2



Average Length



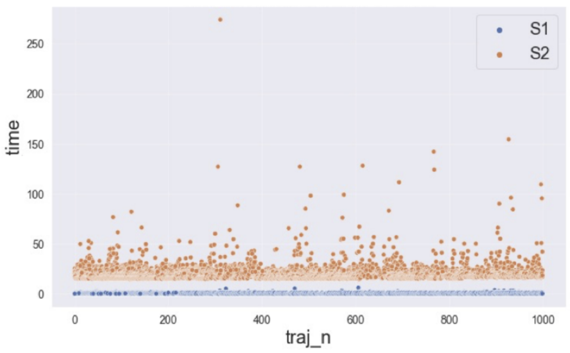
Reward



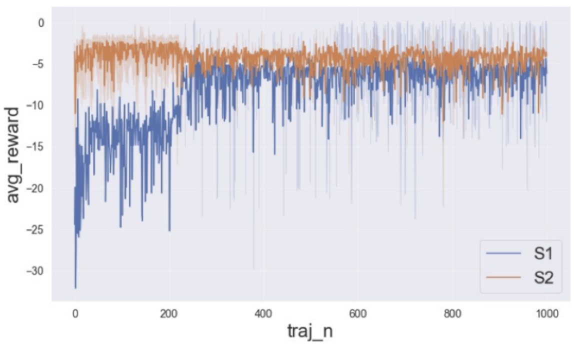
Time

(Aggregated over 10 grids)

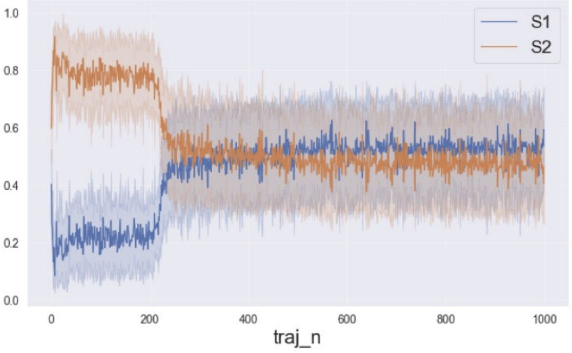
Most interesting case



Time to compute a move



Average Reward for a move

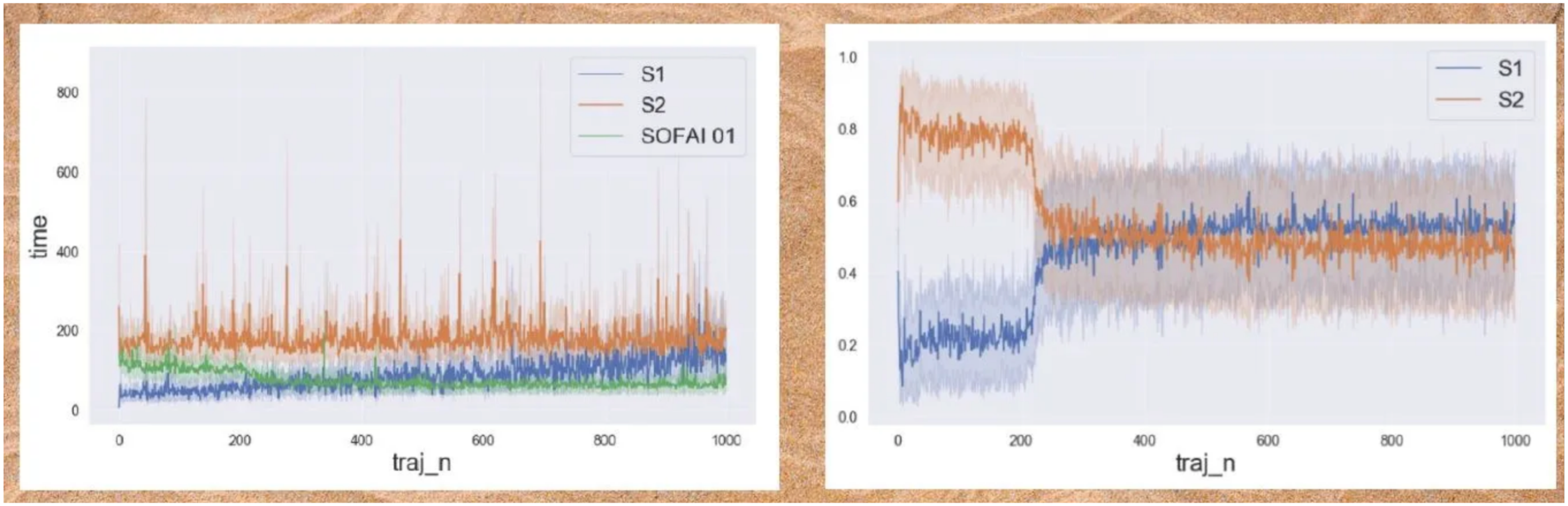


Average fraction of Time each subsystem is used

(Aggregated over 10 grids and 1000 trajectories)

System 1 and System 2 performance

At every step, the SOFAI's meta-cognition unit decides whether it can trust the S1's solution or if it needs to switch to the S2 solver.



Conclusion

- This behavior is similar to what happens in humans
- We first tackle a non-familiar problem with our System 2, until we have enough experience that it becomes familiar and we pass to using System 1