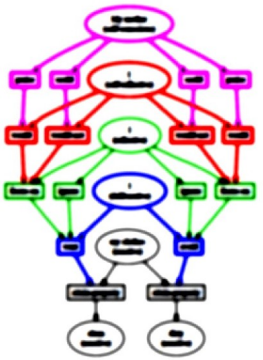
Where Goals Come From

A Model of Self-Conscious Learning

Living Heritage Artificial Intelligence Workshop

June 11th 2009



Example Commonsense Thinking Goals

- I want to pick up that shovel.
- I want to play on the jungle gym.
- I want to show my friend a new trick I discovered.
- I want to avoid being late to school.
- I want to stop my friends from fighting.
- I want to be nice to my sister.
- I want to win the high-jump contest.
- I want to draw a picture of an exciting adventure.

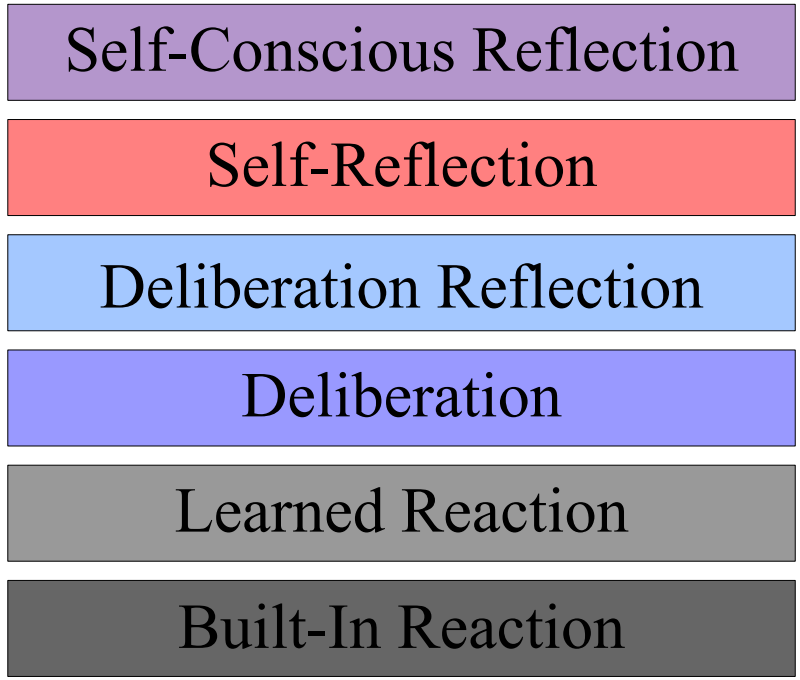
Many Goals are Learned by Failing to Accomplish Goals

- For scooping mud: I should use a spoon next time instead of a fork.
- For kicking a ball: I should be more sure of my footing next time I plan to use my bad knee.
- For getting heavy luggage to the airport: I should consider asking a friend with a car for a ride, so I don't need to struggle on the subway.
- For avoiding offending my hosts: I should let them know I don't eat meat before they cook me dinner.

Minsky's Emotion Machine
Model-6 Layered Theory

Our model is inspired by Minsky's layered "Model-6" theory of reflective thought published in his book *The Emotion Machine*.

An iconic overview is shown here.



Our interpretation considers each layer to be built out of components from the layers below.

Reactive goals are grouped into planned goals.

Ways of deliberating are grouped into reflection goals

Reflection goals are grouped into personalities.

Personality goals form the self-conscious layer.

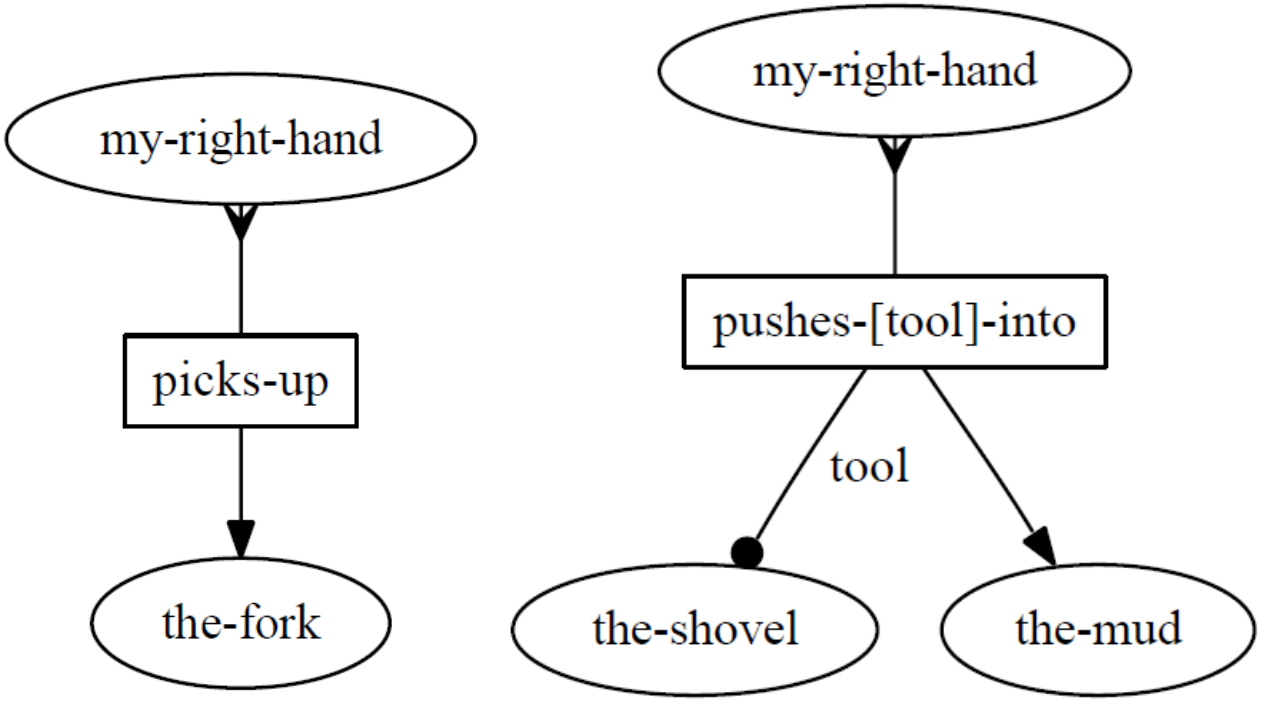
The Story of Muddy Carol

(1) A little girl named Carol is playing alone in the mud. She wants to fill her cup with mud, and first tries to do this with her fork, but this fails because the mud slips through. She succeeds by using her spoon.

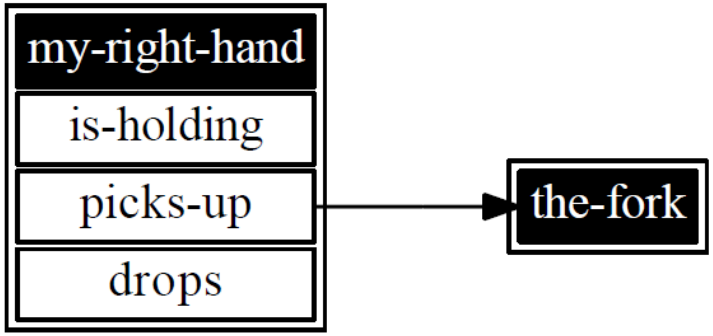
(2) A stranger scolds Carol for playing in the mud. “That is a naughty thing to do.” Carol feels anxious, alarmed, and afraid. Overcome by fear and the urge to escape, she interrupts her present goal and runs to her parent's protection.

(3) Carol returns to her mother for help, but instead of defense or encouragement, all she gets is a reproof, her mother scolds, “What a disgusting mess you've made! See all the mud on your clothes and your face." Carol, ashamed, begins to cry.

Example Semantic Representations



Semantic Graphs



Frame

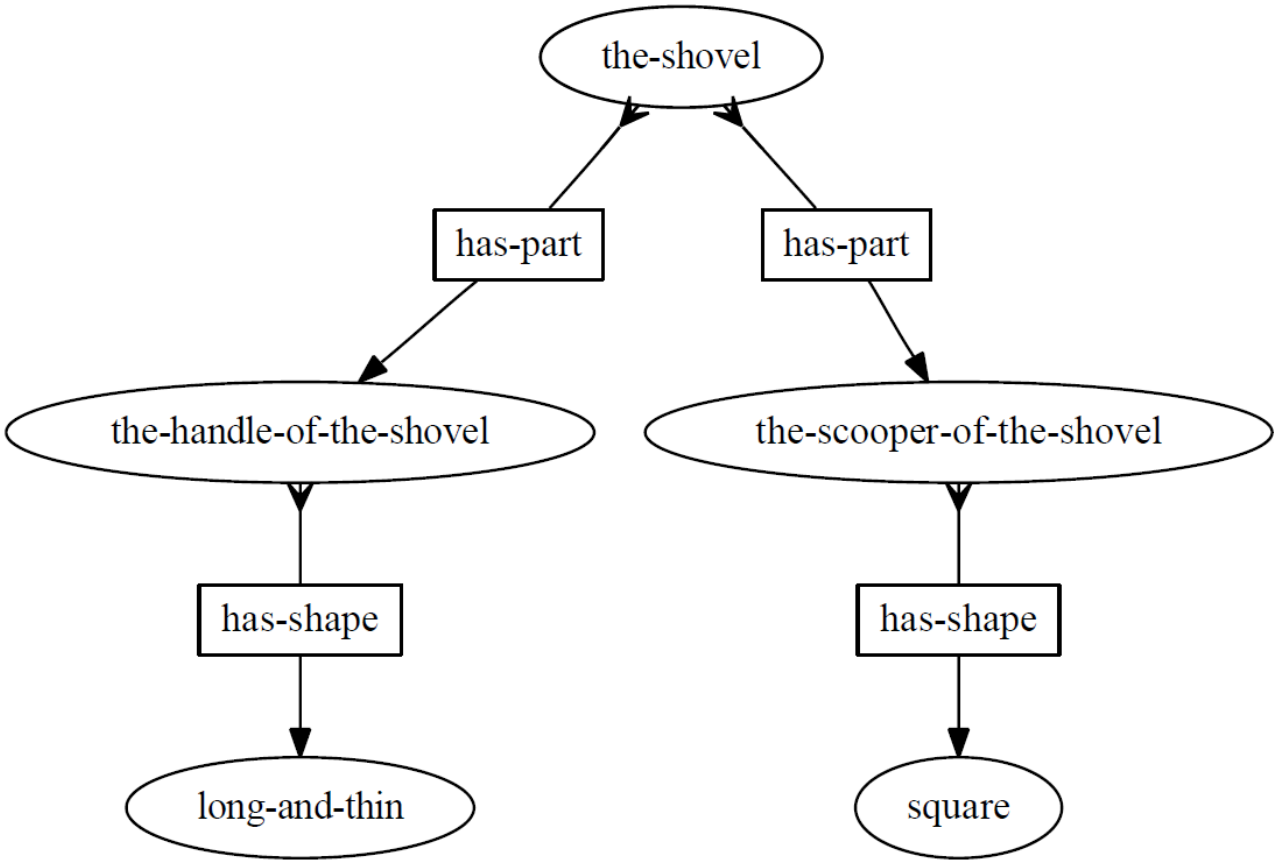
```
picks-up(my-right-hand, the-fork)
pushes-tool-into(my-right-hand,
    the-shovel, the-mud)
```

Logic

```
[picks-up my-right-hand the-fork]
[pushes-tool-into my-right-hand the-shovel the-mud]
```

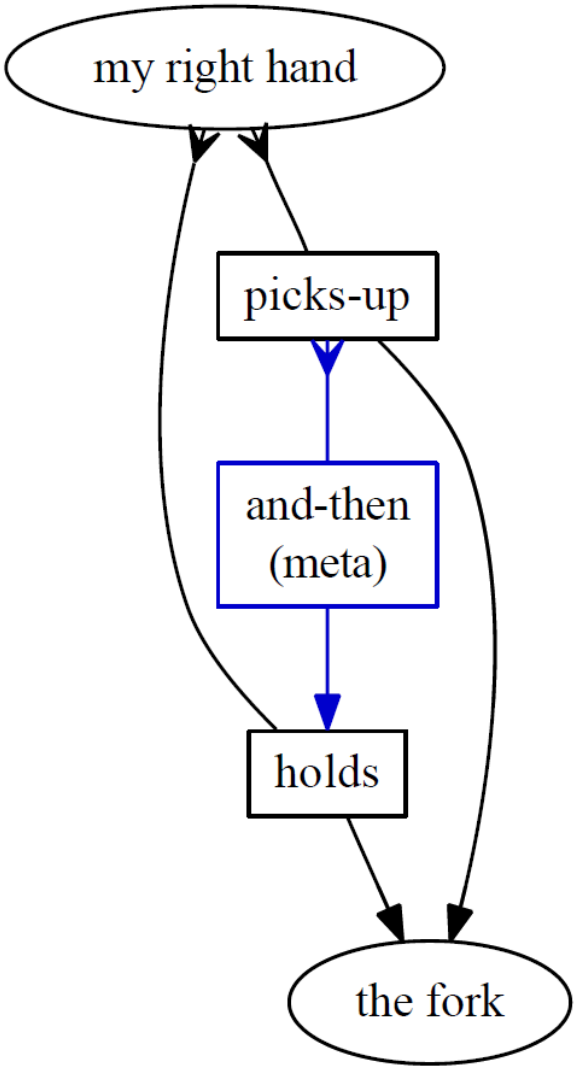
Funk2

Example Physical Object



A simple semantic composition of a physical object as represented by a semantic graph.

Knowledge about Knowledge

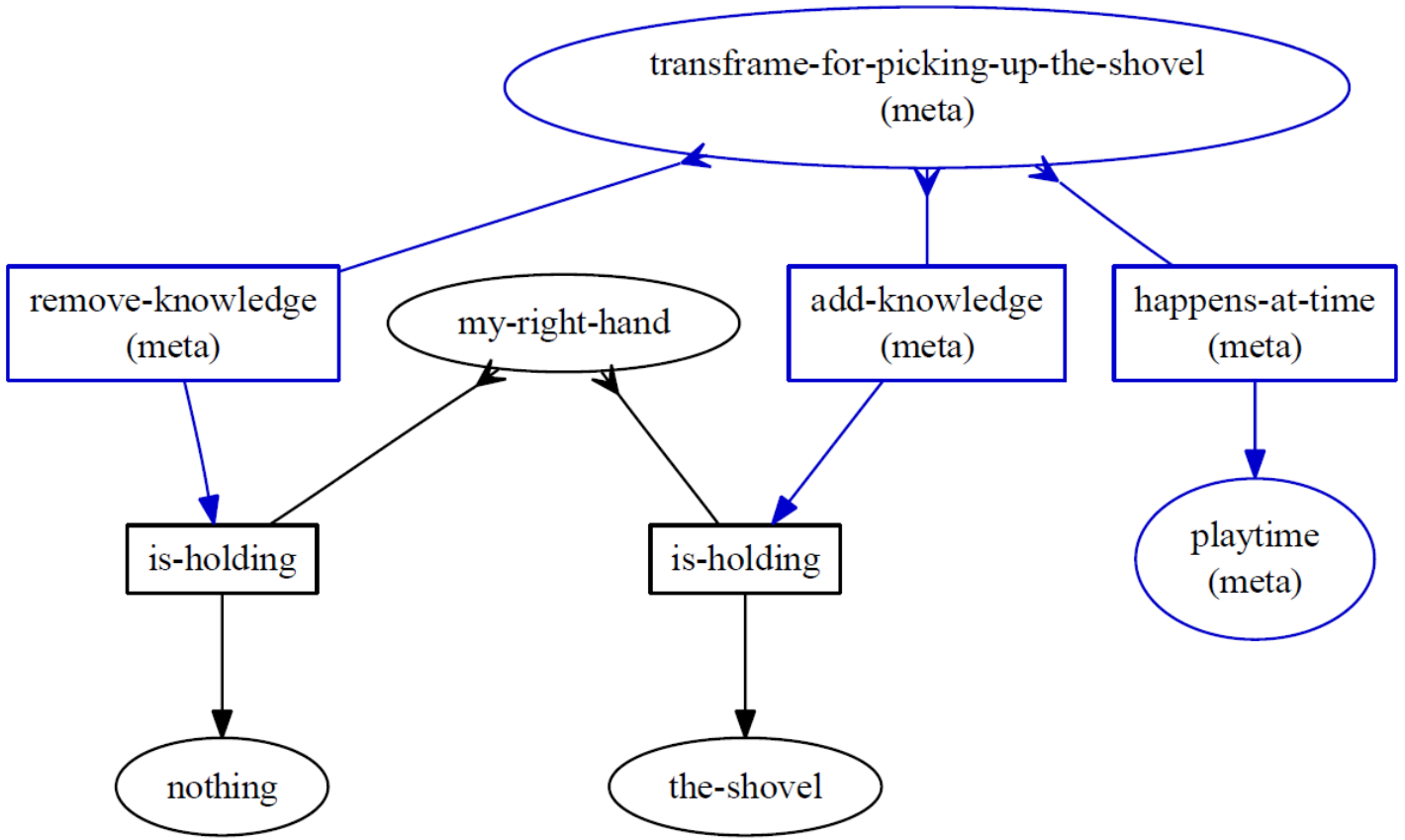


We refer to edges between edges as meta-knowledge or "knowledge about knowledge."

Meta-knowledge is useful for modelling causal relations, temporal relations, and logical relations.

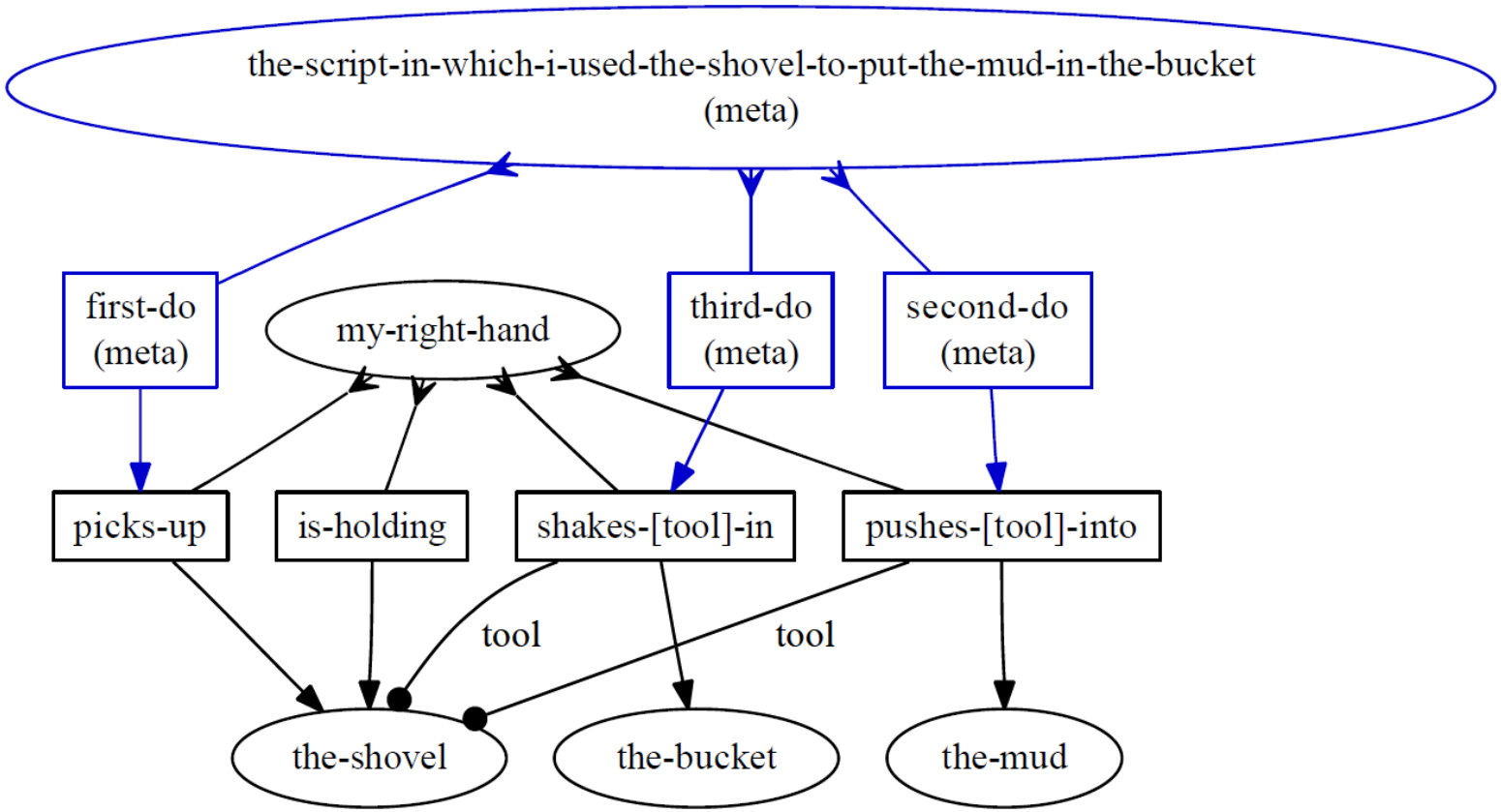
Higher-order relationships are necessary for building reflective layers of focus and control.

Transframe Knowledge



A transframe is a meta-object that can represent changes in knowledge.

Script Knowledge

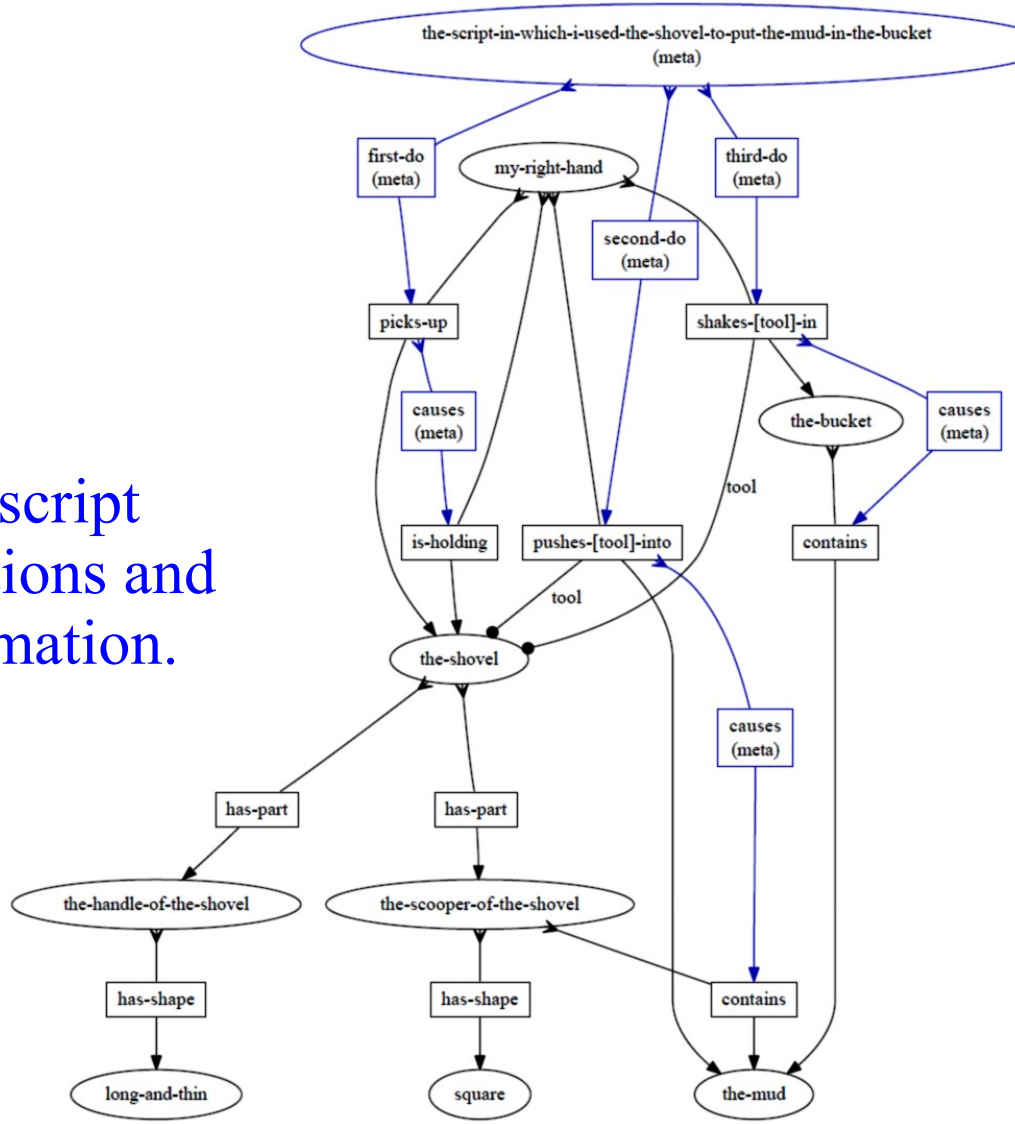


A three step script:

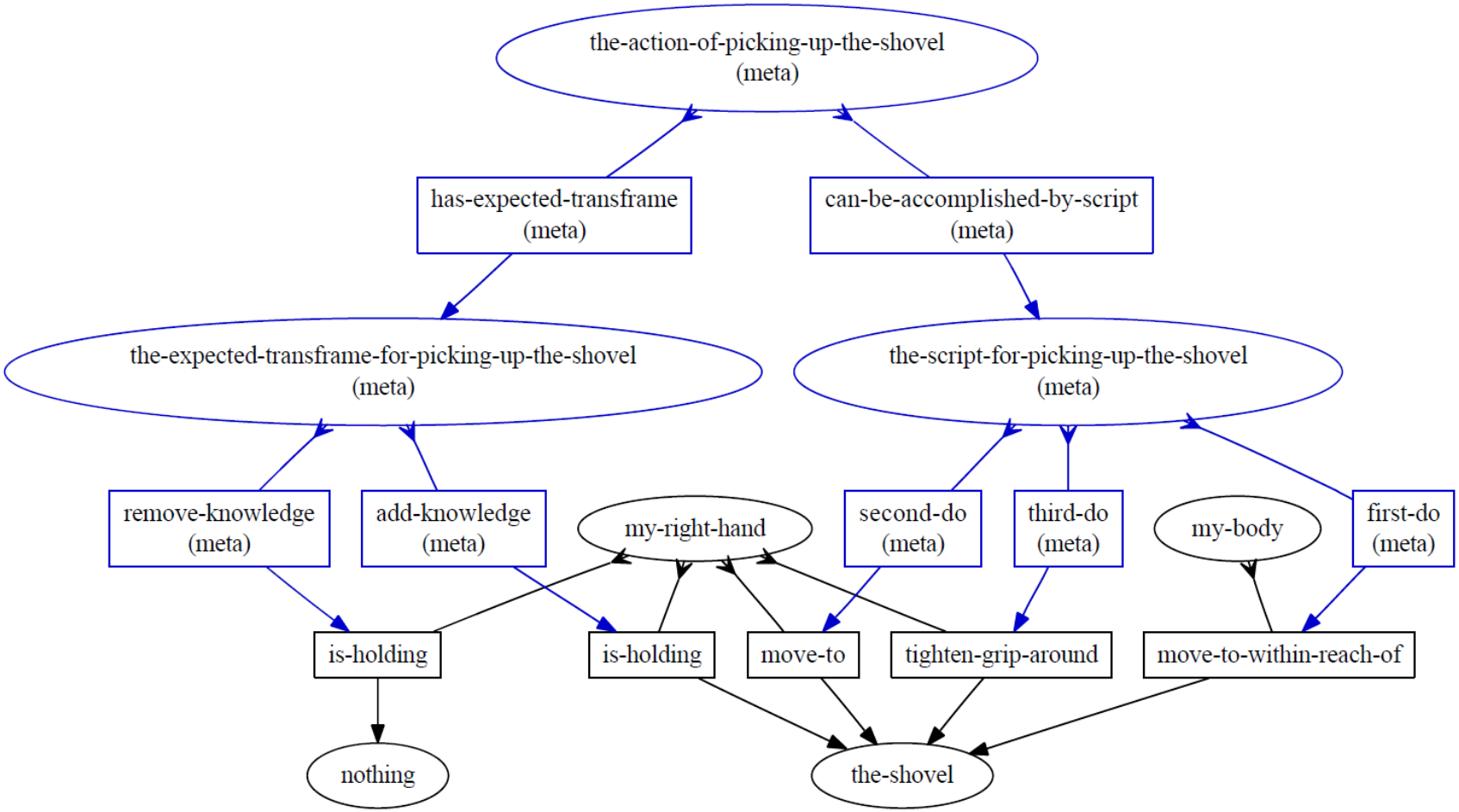
(1) my right hand picks up the shovel,
(2) my right hand pushes the shovel into the mud,
(3) my right hand shakes the shovel in the bucket.

Example Complex Script Knowledge

The same simple three-step script with a few causal meta-relations and other related semantic information.

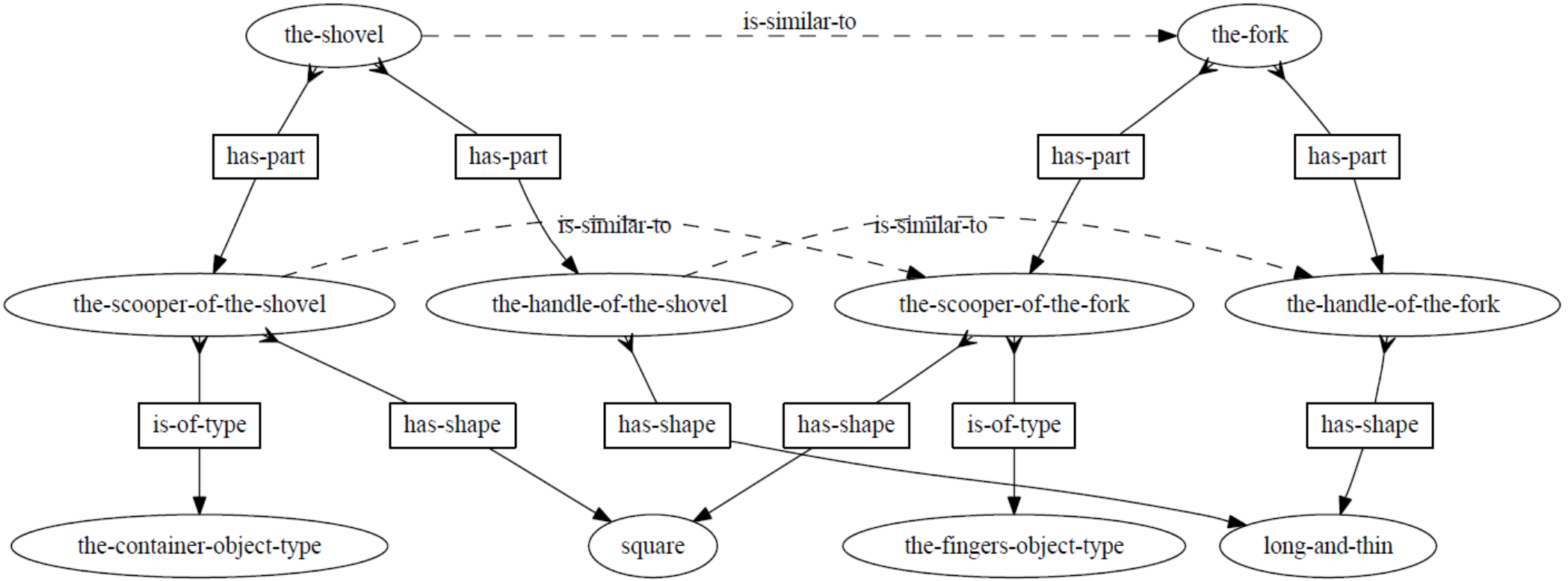


Action Knowledge



Actions consist of a script with an expected transframe.

Example Similarity Mapping

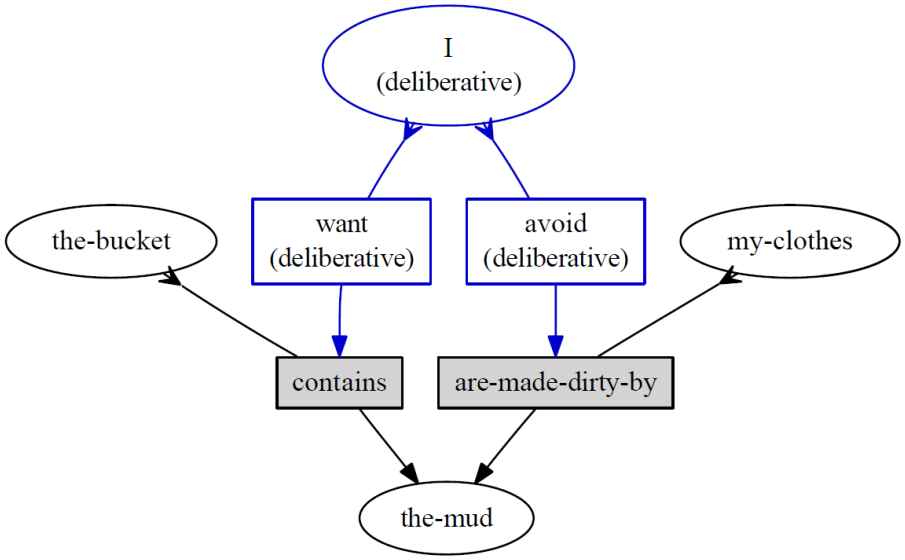


Three points of similarity between a representation of a shovel and a representation of a fork are connected by dotted edges.

Deliberative Knowledge

English Interpretation:

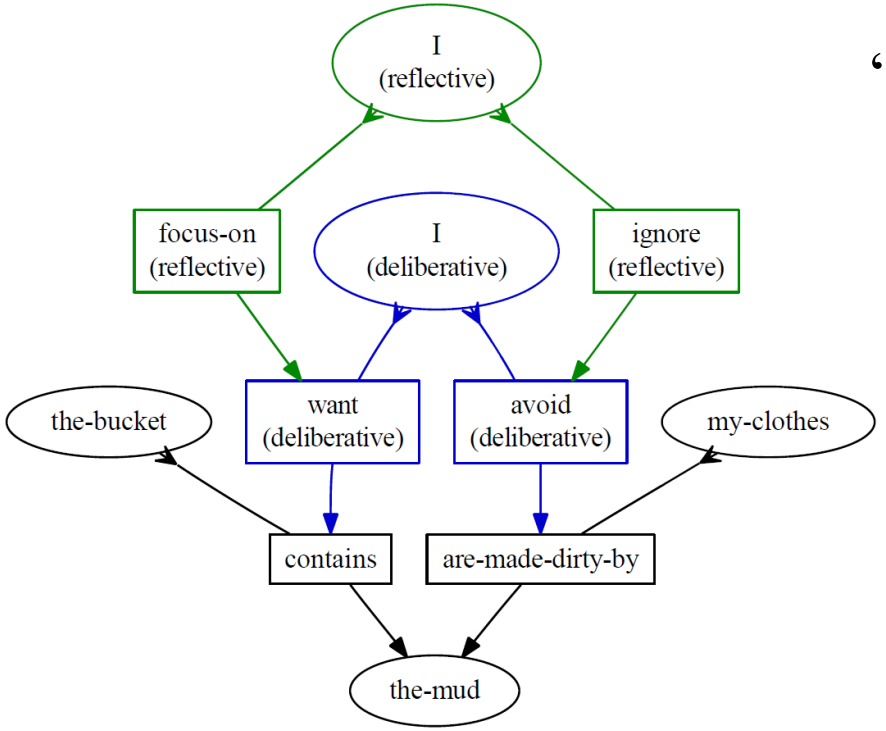
"I avoid
getting my clothes dirty."



Basic goals in our model are represented by what we call "deliberative knowledge" or "the deliberative layer."

Reflective Knowledge

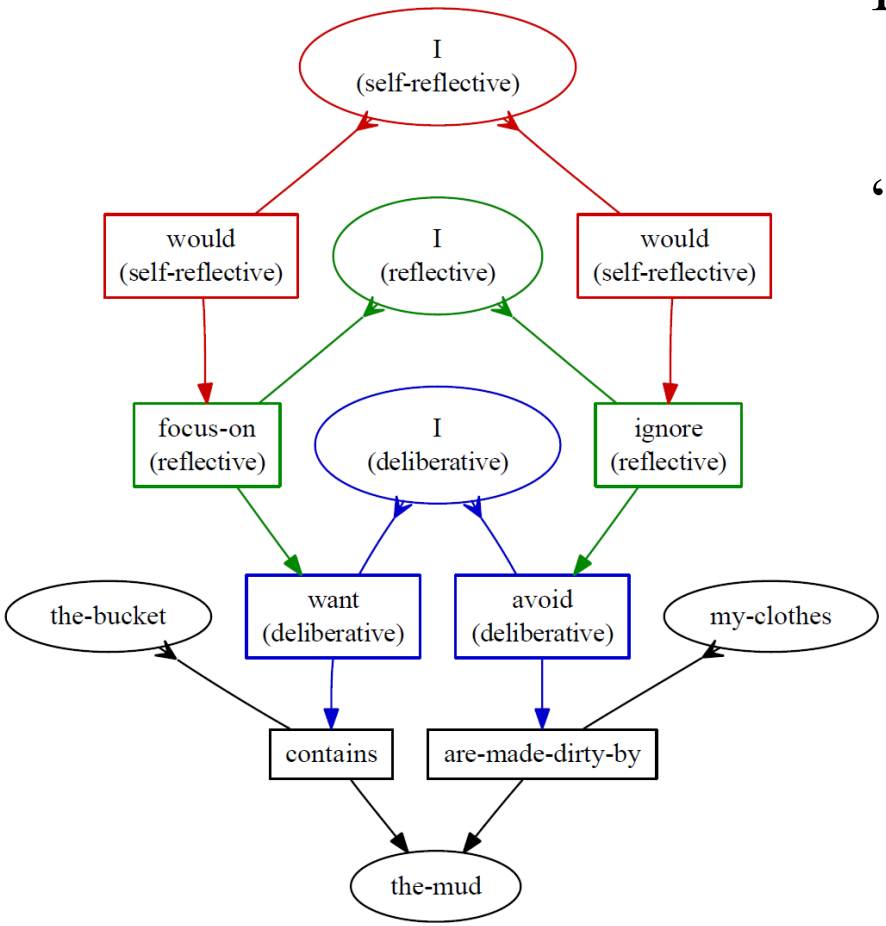
English Interpretation:



"I ignore
avoiding
getting my clothes dirty."

Selecting, constraining, and otherwise reasoning about groups of deliberative goals is handled by what we call "reflective knowledge" or "the reflective layer."

Self-Reflective Knowledge

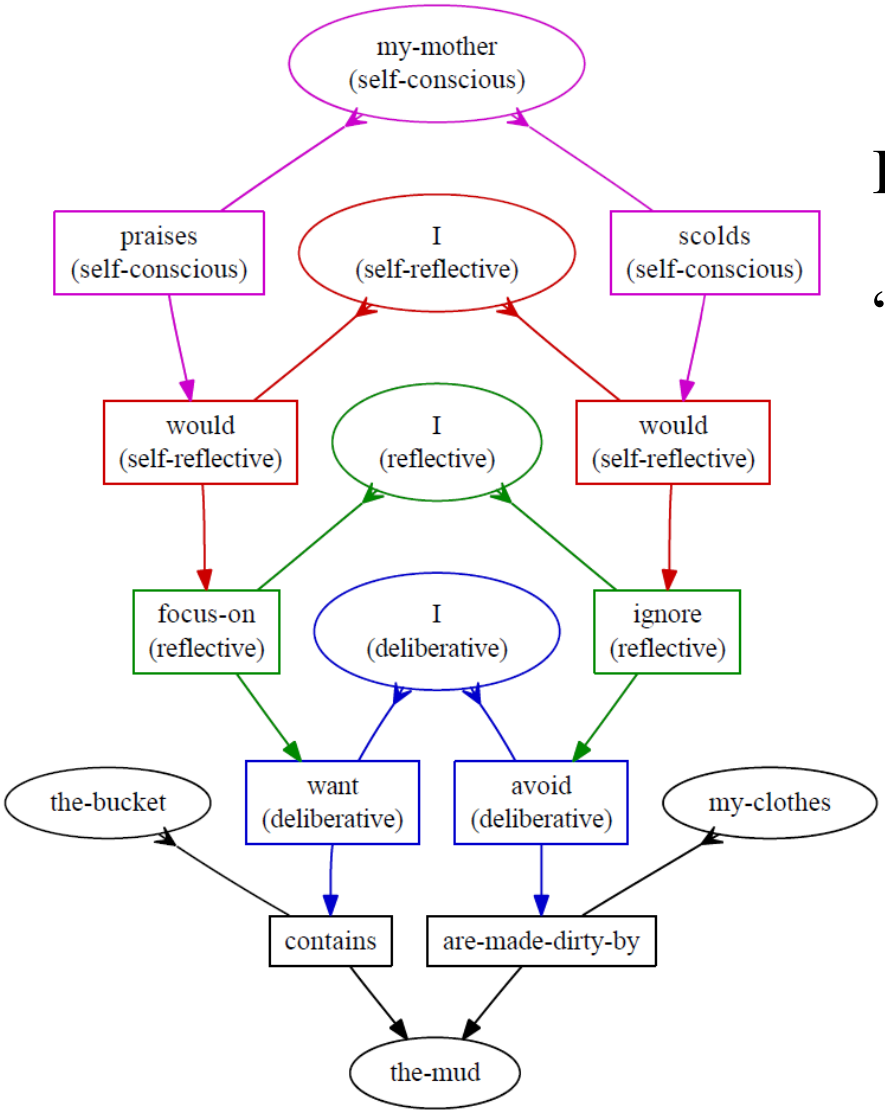


English Interpretation:

"I would
ignore
avoiding
getting my clothes dirty."

Self-models of what one would or would not do (i.e. personality traits) are represented as what we call "self-reflective knowledge" or "the self-reflective layer."

Self-Conscious Knowledge



English Interpretation:

"My mother scolds if
I would
ignore
avoiding
getting my clothes dirty."

Imprimer opinions, inherited personality trait constraints, and other top-level goals are referred to as "self-conscious knowledge," or "the self-conscious layer."

Where Lower-Level Goals Come From

- Our genes determine built-in goals, such as finding food, water, warmth, and love.
- Our experience and deliberation groups sequences of actions into script goals, such as walking quietly, running quickly, acting scary, and resting safely.
- Our reflective thinking develops deliberation goals, such as playing with sticks, daydreaming about the future, considering the causality of failures and successes of the past, and optimizing our planning processes.

Higher-Level Self-Reflective Goals

- Learning about our personalities, leads us to self-reflective goals, such as
  - gathering firewood before playing with sticks,
  - considering safely practicing instead of playing,
  - considering efficient switches between ways of thinking,
  - considering other people's personalities as our own,
  - considering efficient combinations of social or expert ways of thinking,

Highest-Level Self-Conscious Goals

- Learning about the values of different personalities gives us our self-conscious goals, the highest-level goals in our model, such as
  - asking for help from friends is good because it strengthens friendships and is usually more efficient than working alone,
  - practicing is bad because it wastes time and actual performance and getting the job done is what counts, and
  - lying to friends is socially dangerous but can have great benefits as long as they trust me and no one finds out.

Moral Values can be Personality Goals

- “Thou shalt not steal.”
- “Thou shalt not kill.”
- “Do unto others as you would have them do unto you.”
- Obey the law.
- Tell the truth.
- Show self-restraint.

Moral Values can be Personality Goals

- “Thou shalt not steal.”
- “Thou shalt not kill.”
- “Do unto others as you would have them do unto you.”
- Obey the law.
- Tell the truth.
- Show self-restraint.

- Don't be a thief.
- Don't be a murderer.
- Don't be selfish.
- Don't be a criminal.
- Don't be a liar.
- Don't be gluttonous.

So Be Good for Goodness Sake

This work exists thanks to the inspiration, advise, and support of

Marvin Minsky	Society of Mind Group	MIT Media Lab
Joseph Paradiso	Responsive Environments Group	MIT Media Lab
Henry Lieberman	Software Agents Group	MIT Media Lab
Dustin Smith	Software Agents Group	MIT Media Lab
Barbara Barry	Common Sense Group	MIT Media Lab
Gerald Sussman	Neutral Computer Science Research	MIT CS+AI Lab
Michael Cox	IPTO	DARPA

and many others...