Machines must think like humans to build trust

(Harmonizing human and machine rationality)

Professor Peter Bruza
School of Information Systems
QUT
p.bruza@qut.edu.au

Human rationality and Machine rationality

Shared decision making in
environments of high uncertainty

Agenda:

How humans think (make decisions)

How machines think (make decisions)

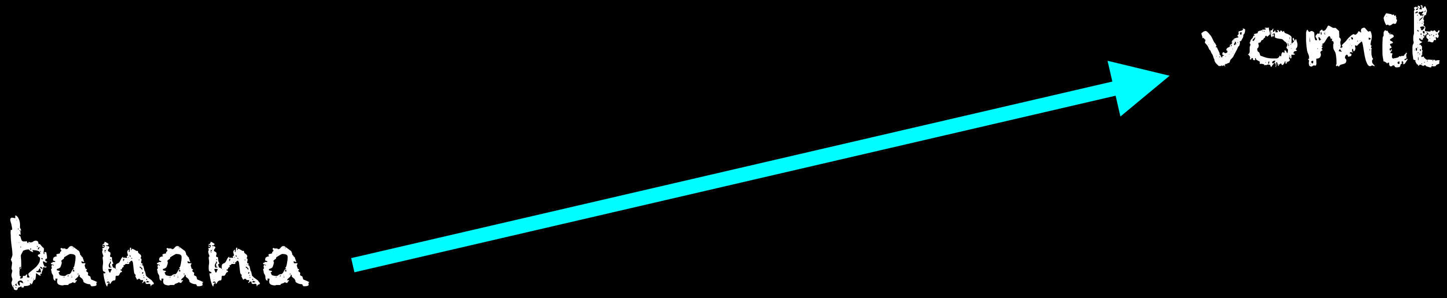
Quantum theory - quantum cognition (to harmonize rationalities)

All: vision for the future

Human thinking (rationality)

Heuristic approach (“bottom up”) - humans learn ad hoc rules heuristics to make decisions; decision performance differs across situations (biases)

Rational approach (“top down”) - people make decisions according to theories based on subjective probability or utility theory; same basic axioms can be used to derive decisions; decision performance is general across situations (axioms) “bottom-up” human thinking...



vomit

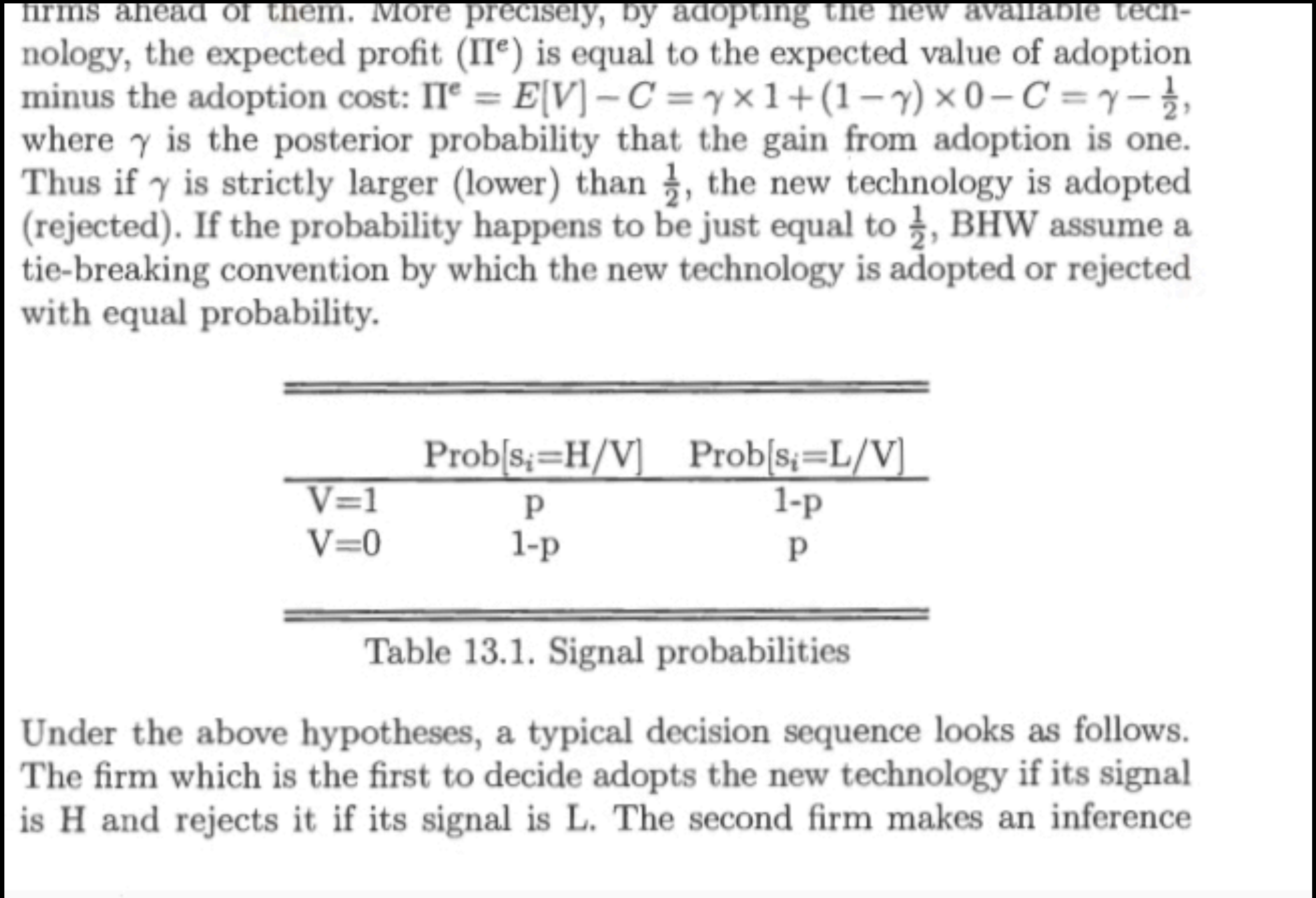
banana

Illusory causation

"At lunchtime, a customer observes all customers that decided before him/her chose restaurant A rather that restaurant B.

He/she may then infer that A is better than B, even if his/her private information implies the opposite".

1) Assumption: agents act rationally (maximize a purely selfish expected utility; their judgements are assumed to be Bayesian - "top-down" explanation)



firms ahead of them. More precisely, by adopting the new available technology, the expected profit (Π^e) is equal to the expected value of adoption minus the adoption cost: $\Pi^e = E[V] - C = \gamma \times 1 + (1-\gamma) \times 0 - C = \gamma - \frac{1}{2}$, where γ is the posterior probability that the gain from adoption is one. Thus if γ is strictly larger (lower) than $\frac{1}{2}$, the new technology is adopted (rejected). If the probability happens to be just equal to $\frac{1}{2}$, BHW assume a tie-breaking convention by which the new technology is adopted or rejected with equal probability.

	Prob[s_i=H/V]	Prob[s_i=L/V]
V=1	p	1-p
V=0	1-p	p

Table 13.1. Signal probabilities

Under the above hypotheses, a typical decision sequence looks as follows. The firm which is the first to decide adopts the new technology if its signal is H and rejects it if its signal is L. The second firm makes an inference

"top down" approaches to human decision making often rely on probability theory

In contrast, potential explanations for "bottom-up" decision making

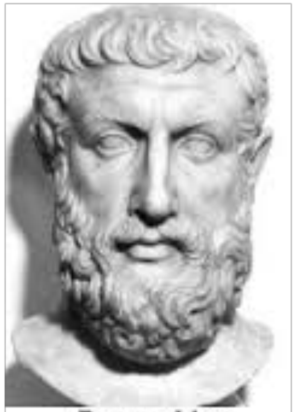
- "follow the herd" (e.g., ad populum)
- "better to be wrong with the majority than right on your own" (conformity bias)

Machine rationality

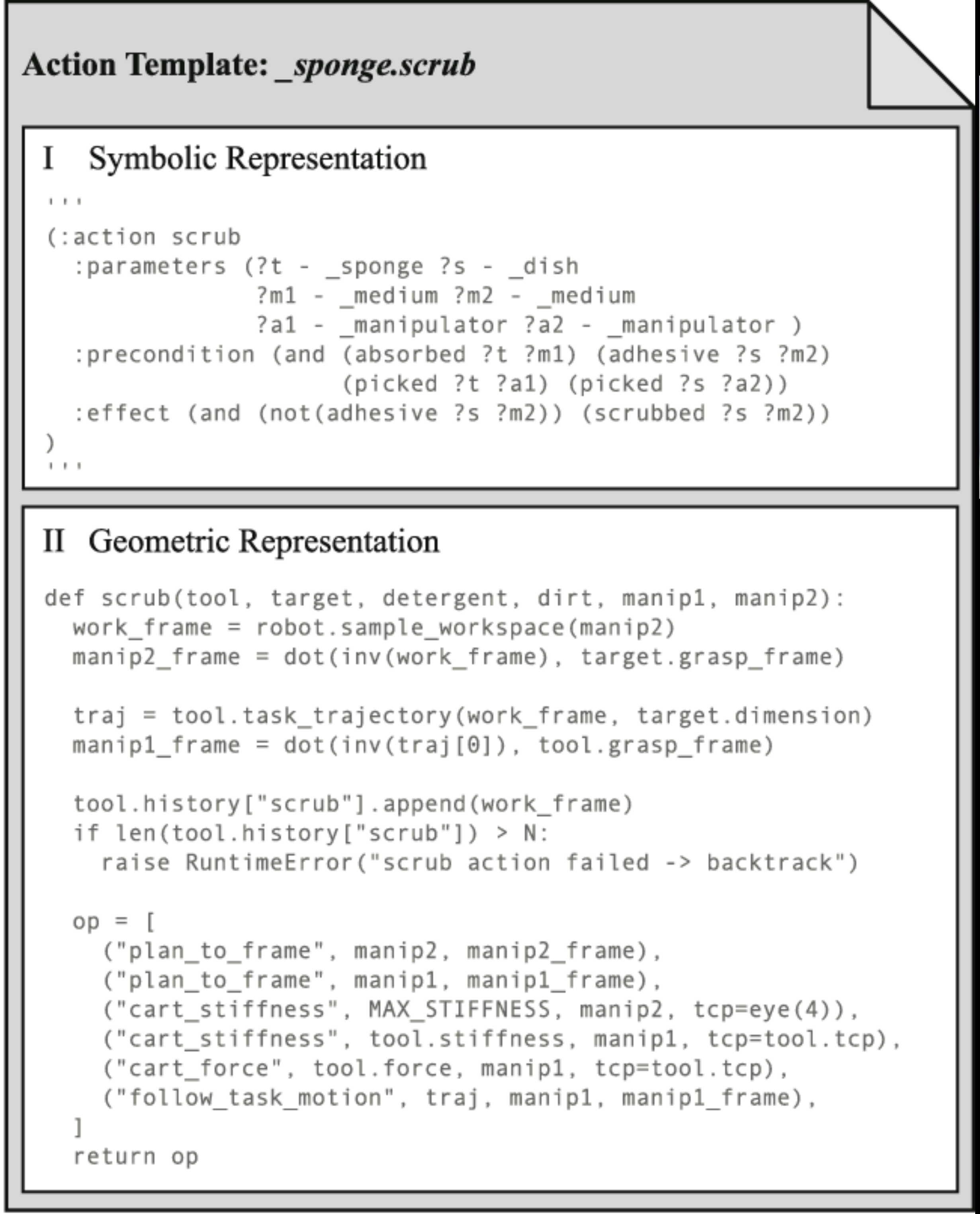
Logical - rationality derives from the laws of some logic

Computational - identifying decisions with highest expected utility, while taking into consideration the costs of computation in complex real-world problems in which most relevant calculations can only be approximated (deep learning)

“It never was
and never will be because it is now,
all together, holding to itself.
For what possible birth of it will you
look for? In what way could it have
grown? From what?
To say or think from “what is not” is
something I won’t allow you,
because there is no saying or
thinking that is not. So it must
either be, completely, or not be”.



Parmenides



Action Template: *_sponge.scrub*

I Symbolic Representation

```
'''
(:action scrub
  :parameters (?t - _sponge ?s - _dish
               ?m1 - _medium ?m2 - _medium
               ?a1 - _manipulator ?a2 - _manipulator )
  :precondition (and (absorbed ?t ?m1) (adhesive ?s ?m2)
                     (picked ?t ?a1) (picked ?s ?a2))
  :effect (and (not(adhesive ?s ?m2)) (scrubbed ?s ?m2))
)
'''
```

II Geometric Representation

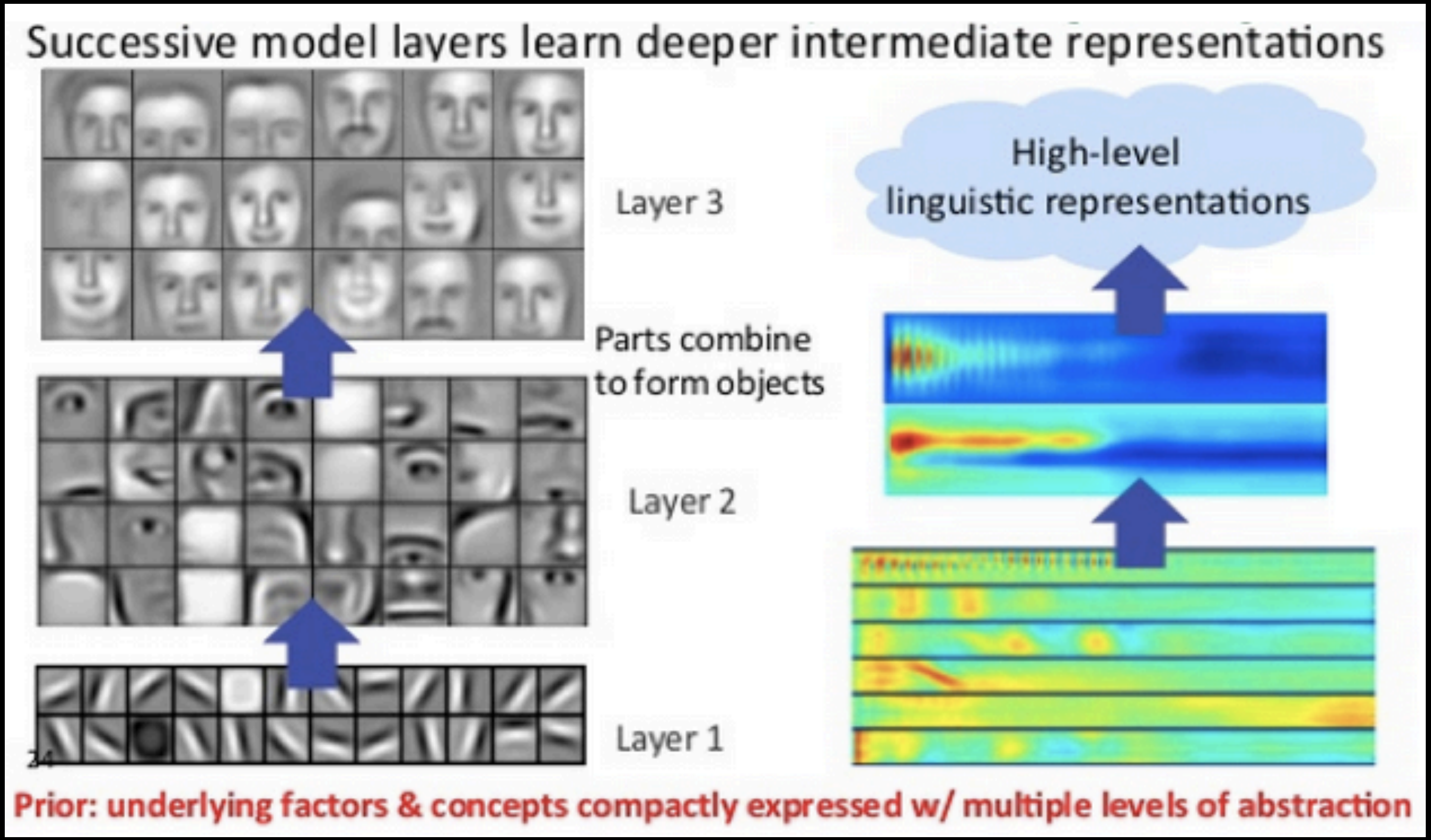
```
def scrub(tool, target, detergent, dirt, manip1, manip2):
  work_frame = robot.sample_workspace(manip2)
  manip2_frame = dot(inv(work_frame), target.grasp_frame)

  traj = tool.task_trajectory(work_frame, target.dimension)
  manip1_frame = dot(inv(traj[0]), tool.grasp_frame)

  tool.history["scrub"].append(work_frame)
  if len(tool.history["scrub"]) > N:
    raise RuntimeError("scrub action failed -> backtrack")

  op = [
    ("plan_to_frame", manip2, manip2_frame),
    ("plan_to_frame", manip1, manip1_frame),
    ("cart_stiffness", MAX_STIFFNESS, manip2, tcp=eye(4)),
    ("cart_stiffness", tool.stiffness, manip1, tcp=tool.tcp),
    ("cart_force", tool.force, manip1, tcp=tool.tcp),
    ("follow_task_motion", traj, manip1, manip1_frame),
  ]
  return op
```

logical (symbolic) approach to machine rationality



Successive model layers learn deeper intermediate representations

Layer 3

Parts combine to form objects

Layer 2

Layer 1

High-level linguistic representations

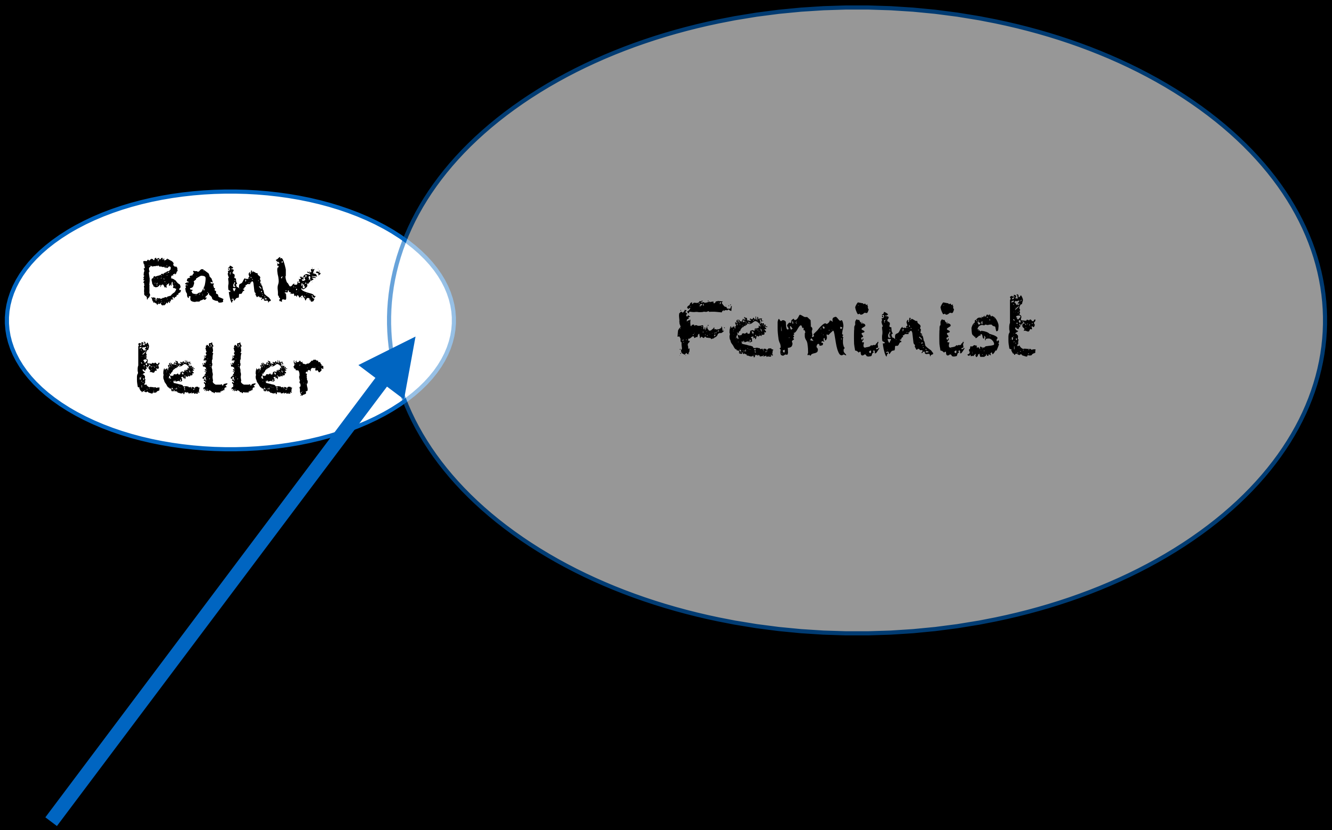
Prior: underlying factors & concepts compactly expressed w/ multiple levels of abstraction

Computational approach to machine rationality (Deep learning)

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations

Which is more probable:

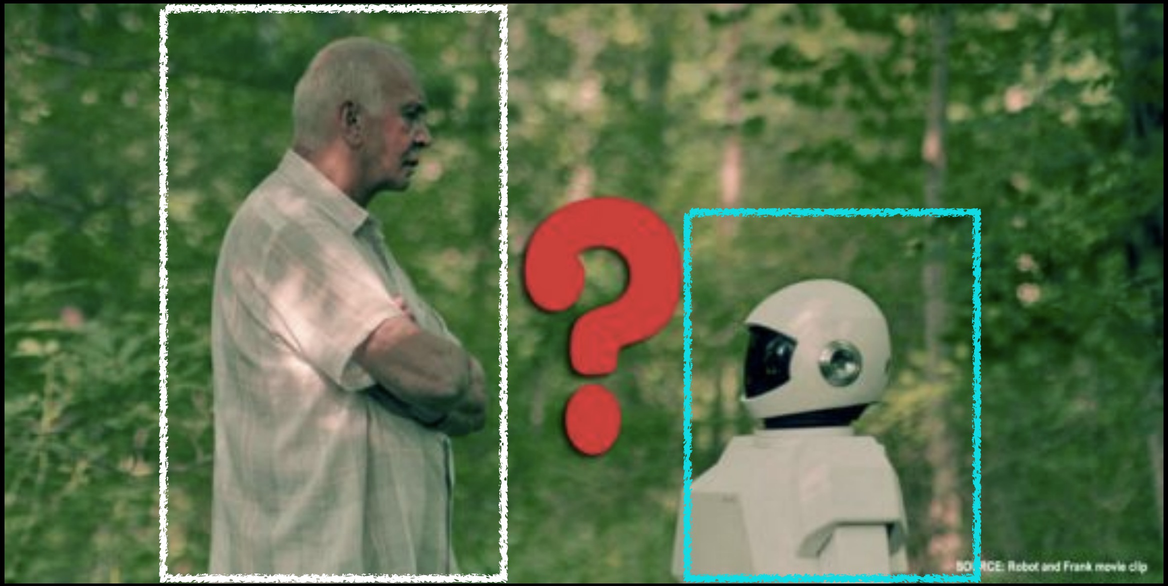
(a)Linda is a bank teller, or
(b)Linda is a bank teller and is active in the feminist movement?



Bank teller AND feminist

So, Linda is a bank teller MUST BE more probable that she is a bank teller AND feminist (.. if we are rational..)

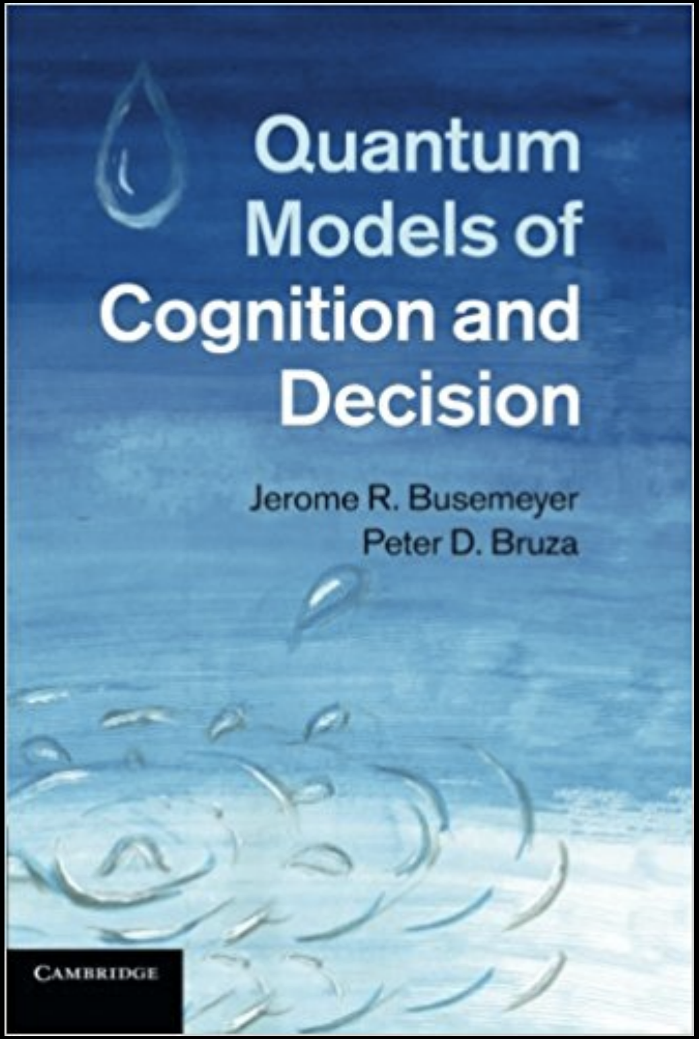
Human rationality vs. Machine rationality



SOURCE: Robot and Frank movie clip

Machine: It's more likely that Linda is just a bank teller (adheres to the law of total probability)

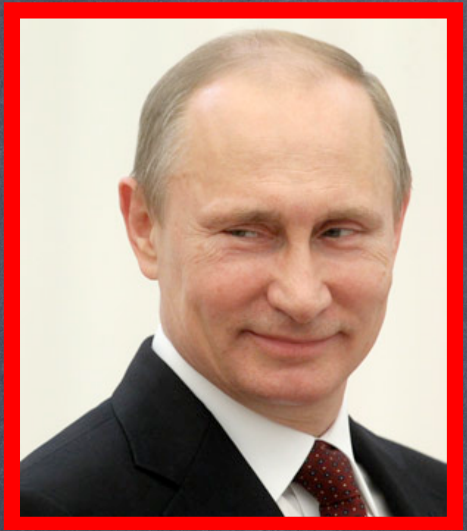
Human: No way! She's a feminist as well! (does NOT adhere to the law of total probability)



Quantum cognition: Like the rational approach to decision making it is based on the axioms of a probability theory (quantum theory).

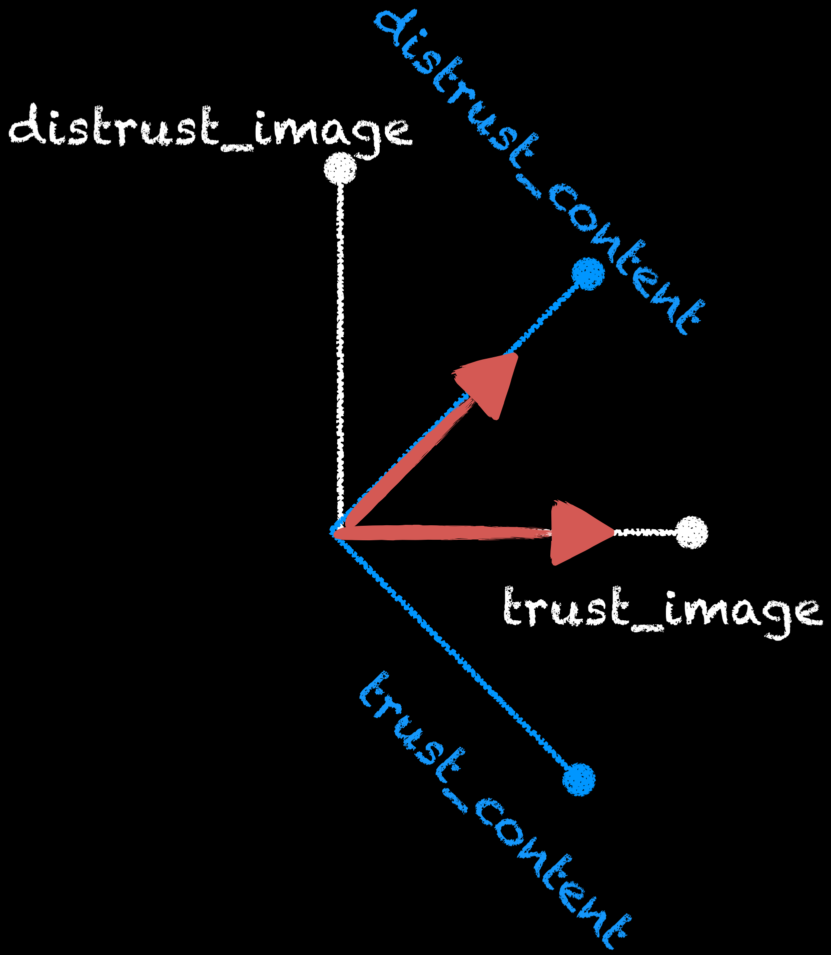
Judgements of image trustworthiness

Does not seem to be photoshopped or altered

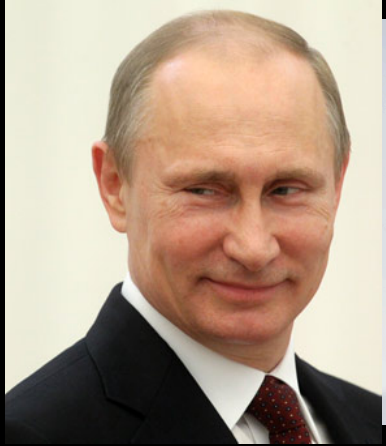


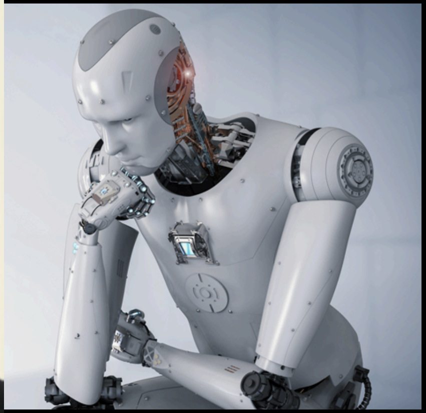
I REALLY COULD NOT SEPARATE WHAT I KNOW ABOUT THIS MAN FROM HIS IMAGE

Do you trust that the image is as an accurate representation of a situation, person or object?



INCOMPATIBLE decision perspectives
(=> law of total probability DOESN'T hold)





trust_ image	trust_ content	prob
y	y	p1
y	n	p2
n	y	p3
n	n	p4

COMPATIBLE decision perspectives
(=> law of total probability DOES hold)

Contextuality

apple chip

How do we ascribe meaning to such novel conceptual combinations?

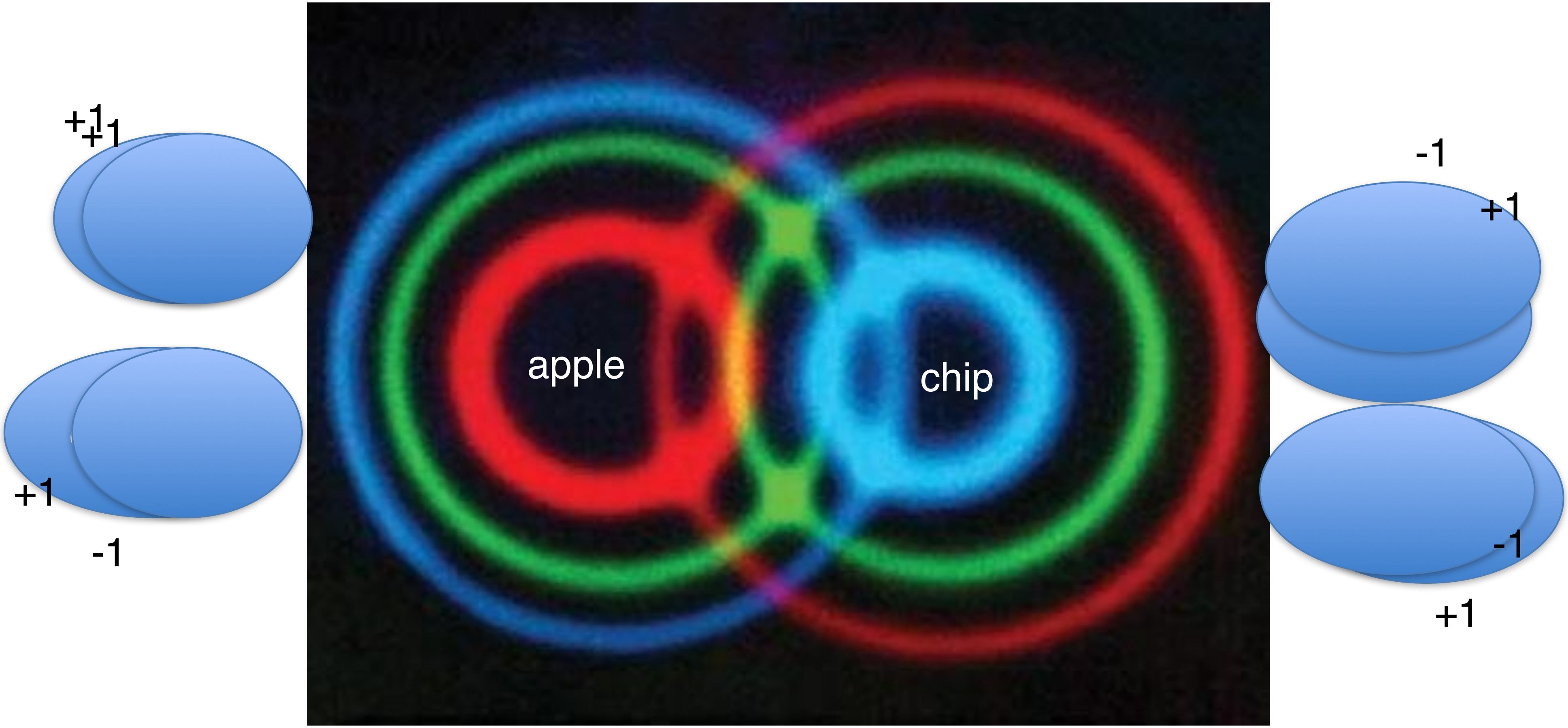
How..? Semantic compositionality

The Principle of Semantic Compositionality (sometimes called 'Frege's Principle') is the principle that the meaning of a (syntactically complex) whole is a function ONLY of the meanings of its (syntactic) parts together with the manner in which these parts were combined. This principle has been extremely influential throughout the history of formal semantics....

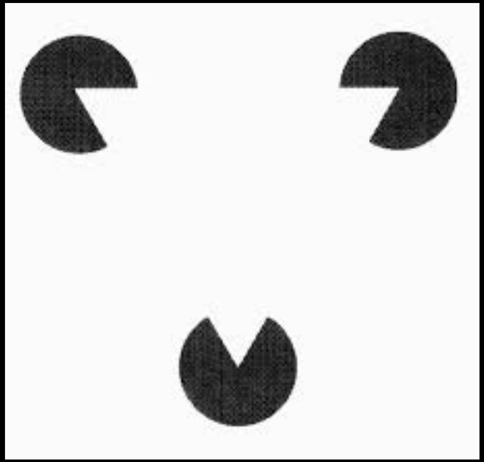
(J. Pelletier, The principle of semantic compositionality, Topoi, vol 13., 1994)

.... in other words WHOLE = SUM of the PARTS

APPLE CHIP = APPLE + CHIP



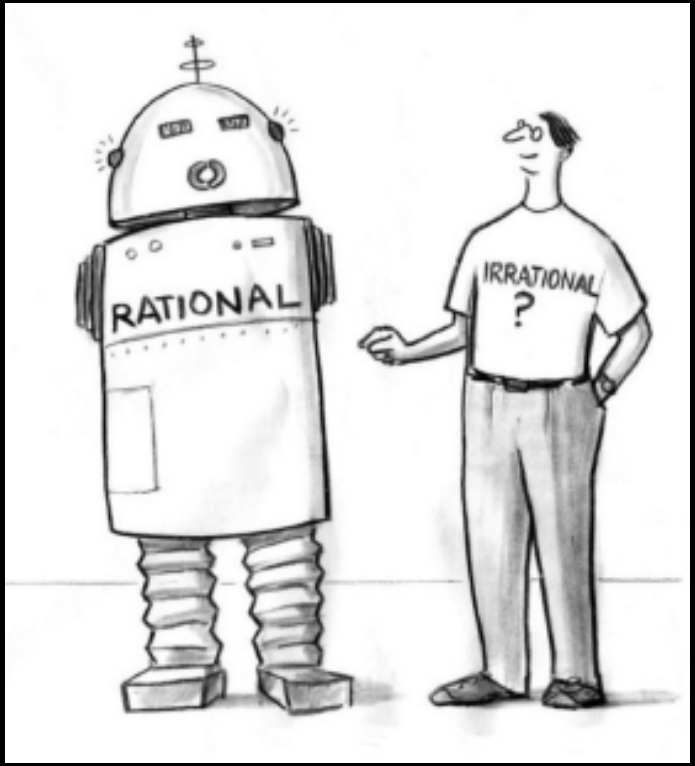
“a nano-chipped granny smith”
“dried pieces of apple that you eat”
N.B. the primes set the context, but do not determine the outcome



Quantum contextuality rules out “whole = sum of the parts” thinking

Logical rationality is necessarily non-contextual as it is based on the principle of semantic compositionality

How can we combine Machine rationality and human rationality in a principled way?



Use the formalism of quantum physics.
A principled framework to model human "irrationality".
Embed this in machines so they "understand" where humans are coming from

Communal visioning exercise

Human rationality and Machine rationality

Shared decision making in environments of high uncertainty

- If we had a magic wand, what would "good" human-machine shared decision making look like?

- How would humans and machines be collaborating in their shared decision making? What would the nature of their interactions be?

- What are we assuming (that might be holding us back)? How might we think about it VERY differently?

- What's ONE exciting unexpected thing that has been uncovered?

Lauren (RA)
[psychologist]

Abdul (PhD)
[theoretical computer science]

Dr. Shahram (Postdoc)
[quantum physicist]

Prof. J. Busemeyer
Indiana Uni.
[cognitive decision theorist]

Dr. P. Wittek
Uni. Toronto
[quantum ML]