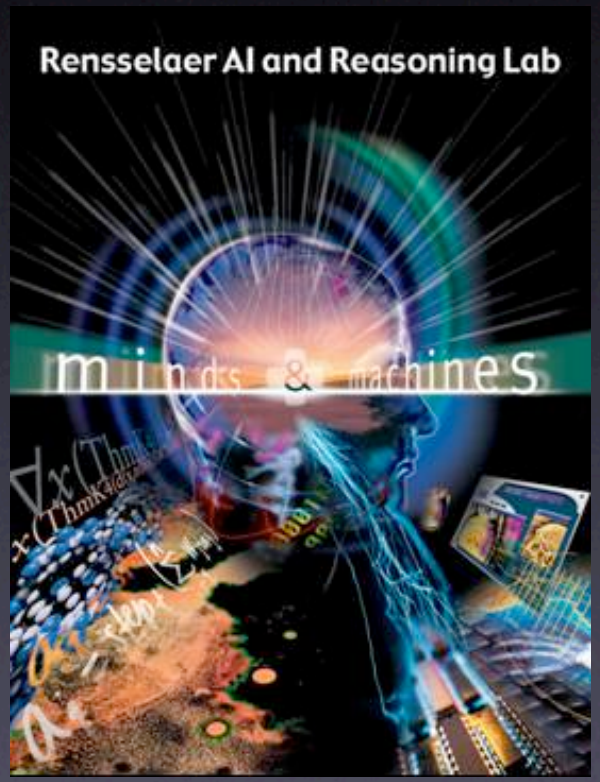
Thoughts on:
Robotics, Free Will, and Predestination



Selmer Bringsjord

(with help from Bettina Schimanski)
Rensselaer AI & Reasoning (RAIR) Laboratory
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 US
CAP@OSU 8.5.05

Rensselaer DEPARTMENT OF COGNITIVE SCIENCE

Rensselaer Computer Science

Are AI and Philosophy of Religion intertwined?

Are AI and Philosophy of Religion intertwined?

Hitherto, probably not.

Let's change that.

T: God exists.

A: But there is evil in the world.

T: Yes, there is. Some free agents decide to do some evil things.

A: But there is *natural* evil in the world.

T: Yes, there is. Some *very powerful* free agents do some evil things.

A: But God could've created a world populated by interacting free agents that don't go wrong.

Maybe, maybe not.

A Relevant Argument to Explore

If God could've created a world populated by interacting free agents that don't go wrong, then AI engineers can build a microworld populated by interacting free robots that don't go wrong.

AI engineers can't build such a microworld.

Let's try to falsify

Therefore by *modus tollens*:

It's not the case that God could've done that.

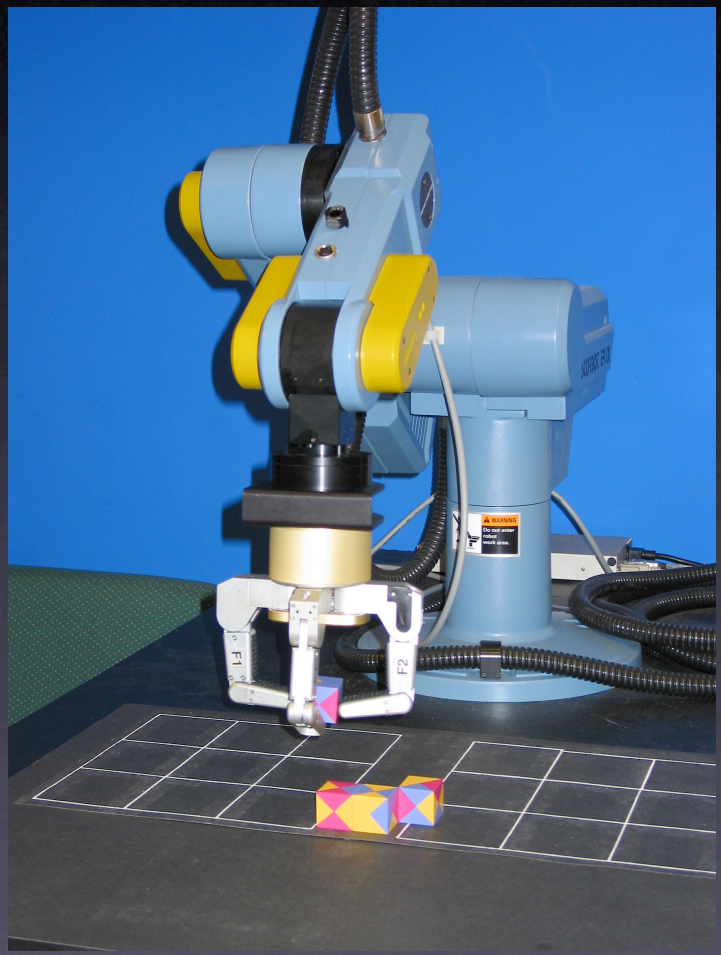
And we'll start by targeting an even simpler engineering goal:

A microworld in which *one* robot — PERI — freely performs *one* morally permissible action.

PERI

Pscyhometric Experimental Robotic Intelligence

- Scorbot-ER IX
- Sony B&W XC55 Video Camera
- Cognex MVS-8100M Frame Grabber
- Dragon Naturally Speaking Software
- NL (Carmel & RealPro?)
- BH8-260 BarrettHand Dexterous 3-Finger Grasper System



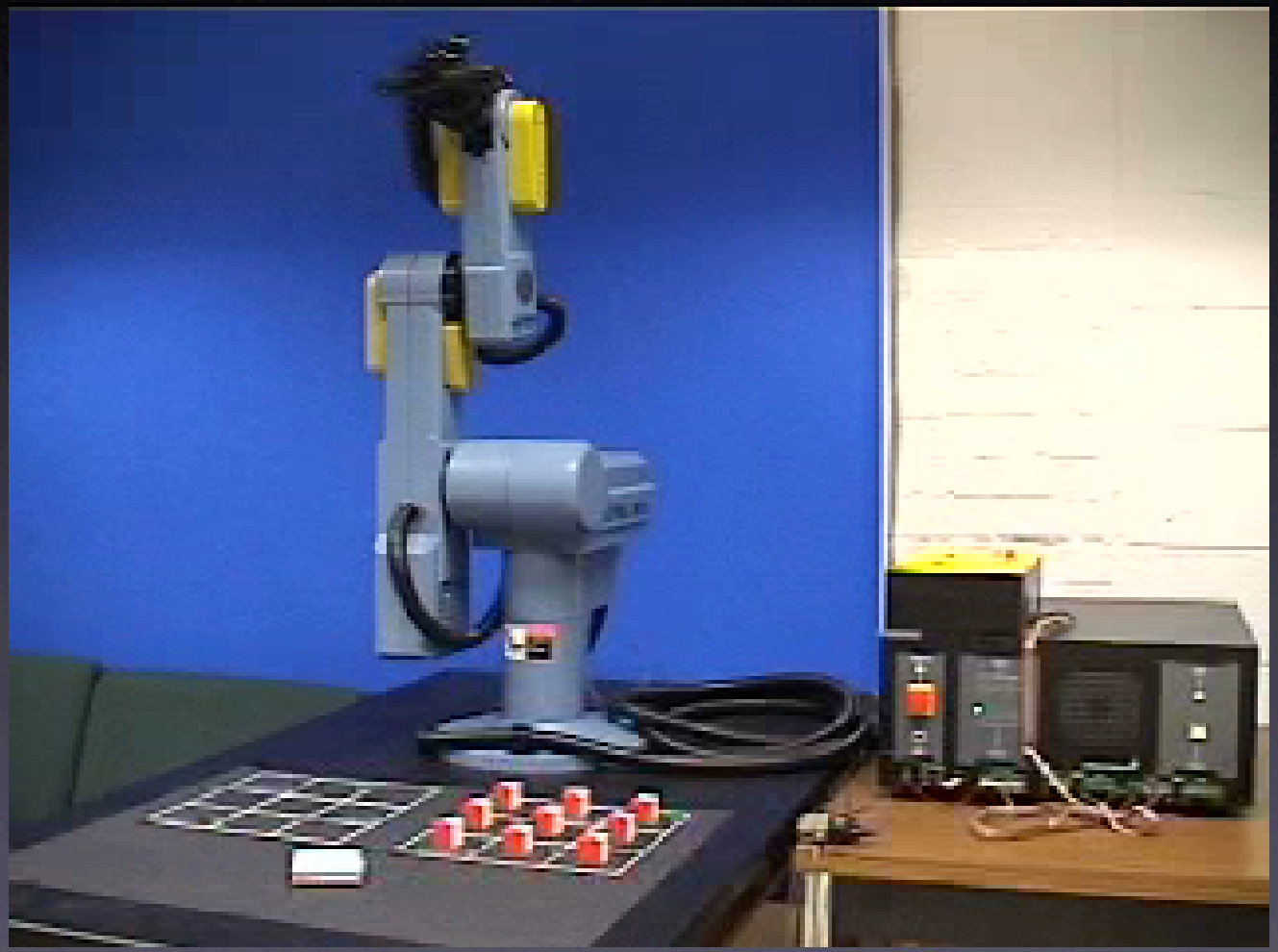
Dear Paul (CEA, Uppsala),

Since I can see color by virtue of a computational implementation of your color perception theory, and yet am no more conscious than a greasy metal gear assembly you have made no progress toward *explaining* (rather than merely *predicting*) qualia.

Sincerely,

PERI

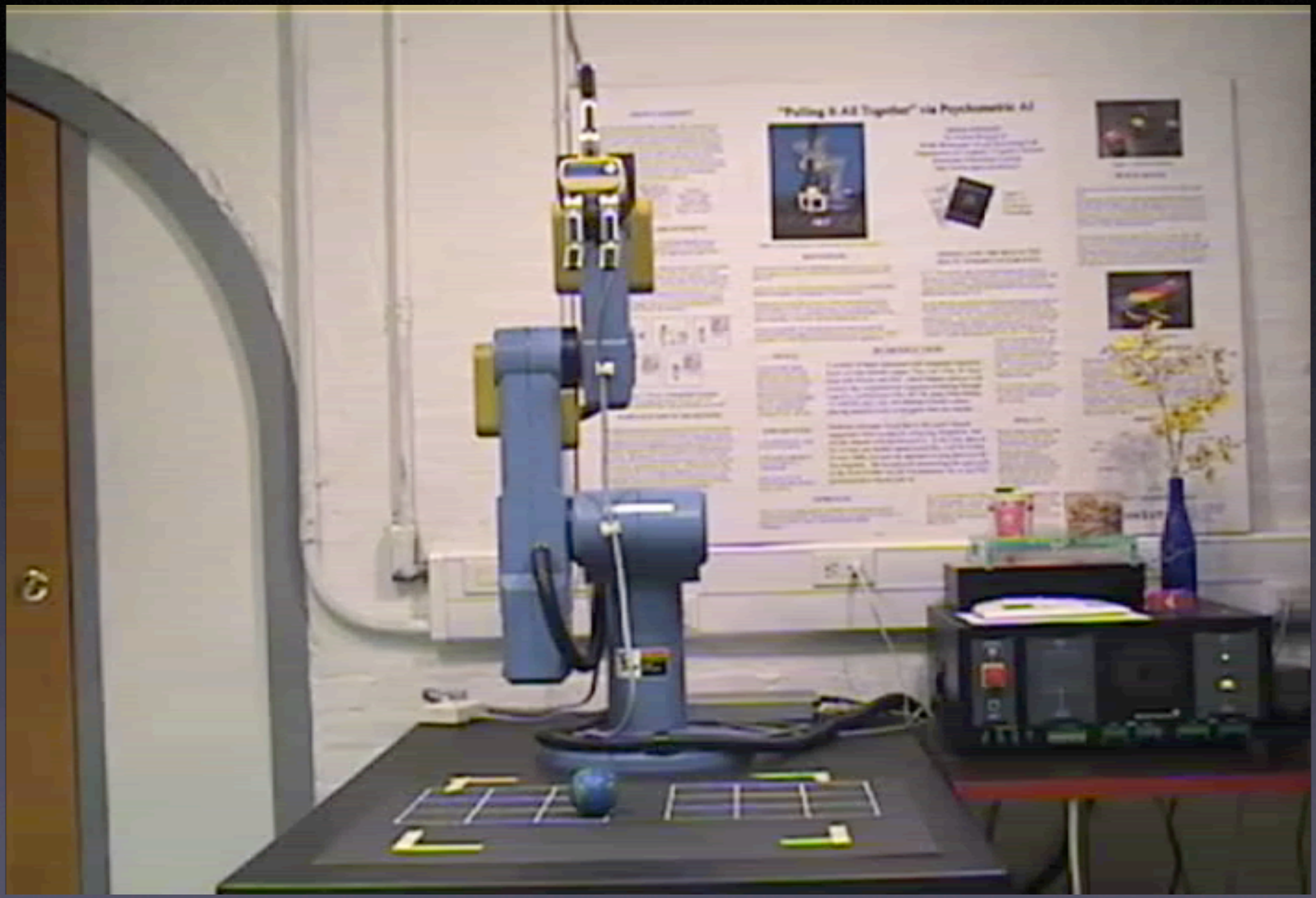
http://www.cogsci.rpi.edu/research/rair/pai



First, a human person freely performing one morally permissible action:

Selmer ...

PERI prepares: `(pickup-earth)` fired...



Option #1 for PERI...

J. EXPT. THEOR. ARTIF. INTELL. 12(2000)341–352

Free will—even for robots

JOHN MCCARTHY

Computer Science Department, Stanford University,
Stanford, CA 94305, USA
e-mail: jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

Abstract. Human free will is a product of evolution and contributes to the success of the human animal. Useful robots will also require free will of a similar kind, and we will have to design it into them.

Free will is not an all-or-nothing thing. Some agents have more free will, or free will of different kinds, than others, and we will try to analyse this phenomenon. Our objectives are primarily technological, i.e. to study what aspects of free will can make robots more useful, and we will not try to console those who find determinism distressing. We distinguish between having choices and being conscious of these choices; both are important, even for robots, and consciousness of choices requires more structure in the agent than just having choices and is important for robots. Consciousness of free will is therefore not just an epiphenomenon of structure serving other purposes.

Free will does not require a very complex system. Young children and rather simple computer systems can represent internally *'I can, but I won't'* and behave accordingly.

Naturally I hope this detailed *design stance* will help understand human free will. It takes the *compatibilist* philosophical position.

There may be some readers interested in what the paper says about human free will and who are put off by logical formulas. The formulas are not important for the arguments about human free will; they are present for people contemplating AI systems using mathematical logic. They can skip the formulas, but the coherence of what remains is not absolutely guaranteed.

Keywords: free will, robots, agents

I can, but I won't.[1]

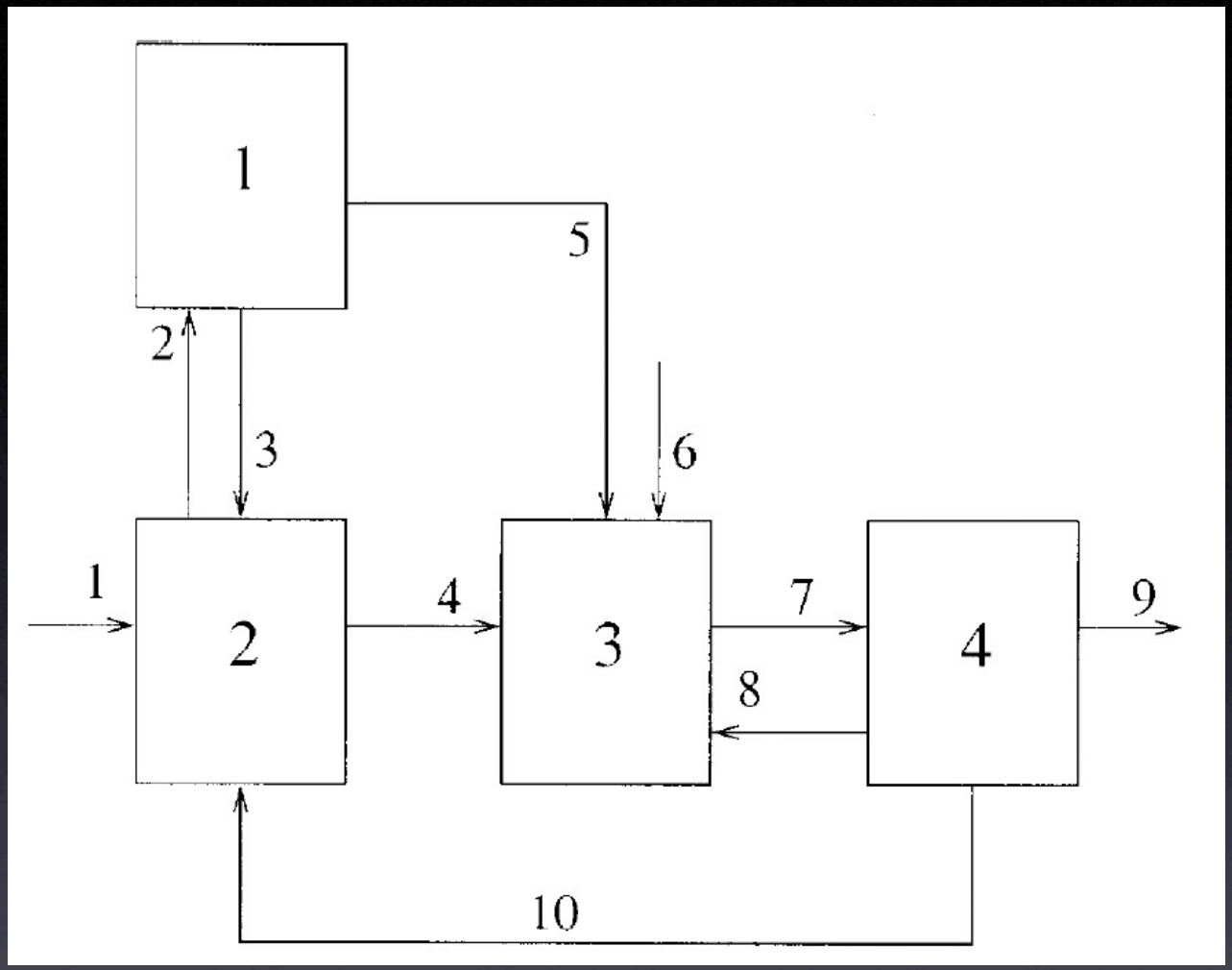
1. Introduction—two aspects of free will

Free will, both in humans and in computer programs has two aspects—the *external* aspect and the *introspective* aspect.

The external aspect is the set of results that an agent P can achieve, i.e. what it *can* do in a situation s:

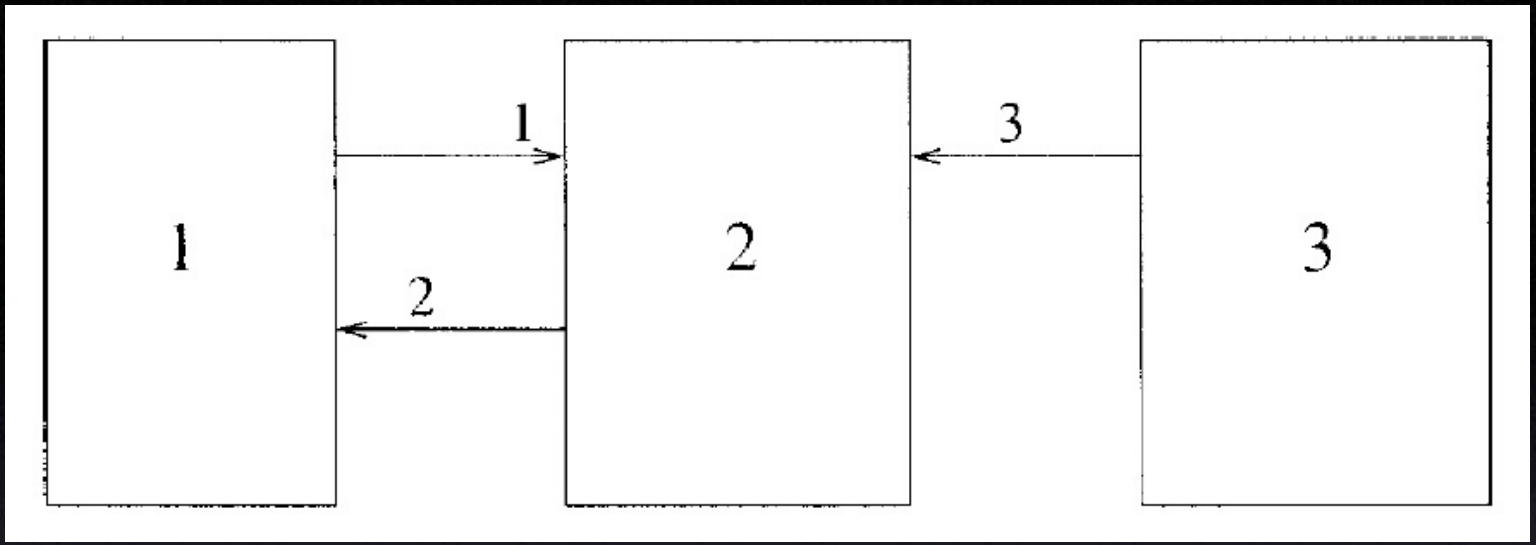
$$Poss(P,s) = \{x | Can(P,x,s)\}. \quad (1)$$

Journal of Experimental & Theoretical Artificial Intelligence ISSN 0952-813X print/ISSN 1362-3079 online © 2000 Taylor & Francis Ltd
http://www.tandfco.uk/journals



$$s_2(t) = S_2(a_2(t))$$
$$s_3(t) = S_3(a_1(t))$$
$$s_4(t) = S_4(a_2(t))$$
$$\vdots$$

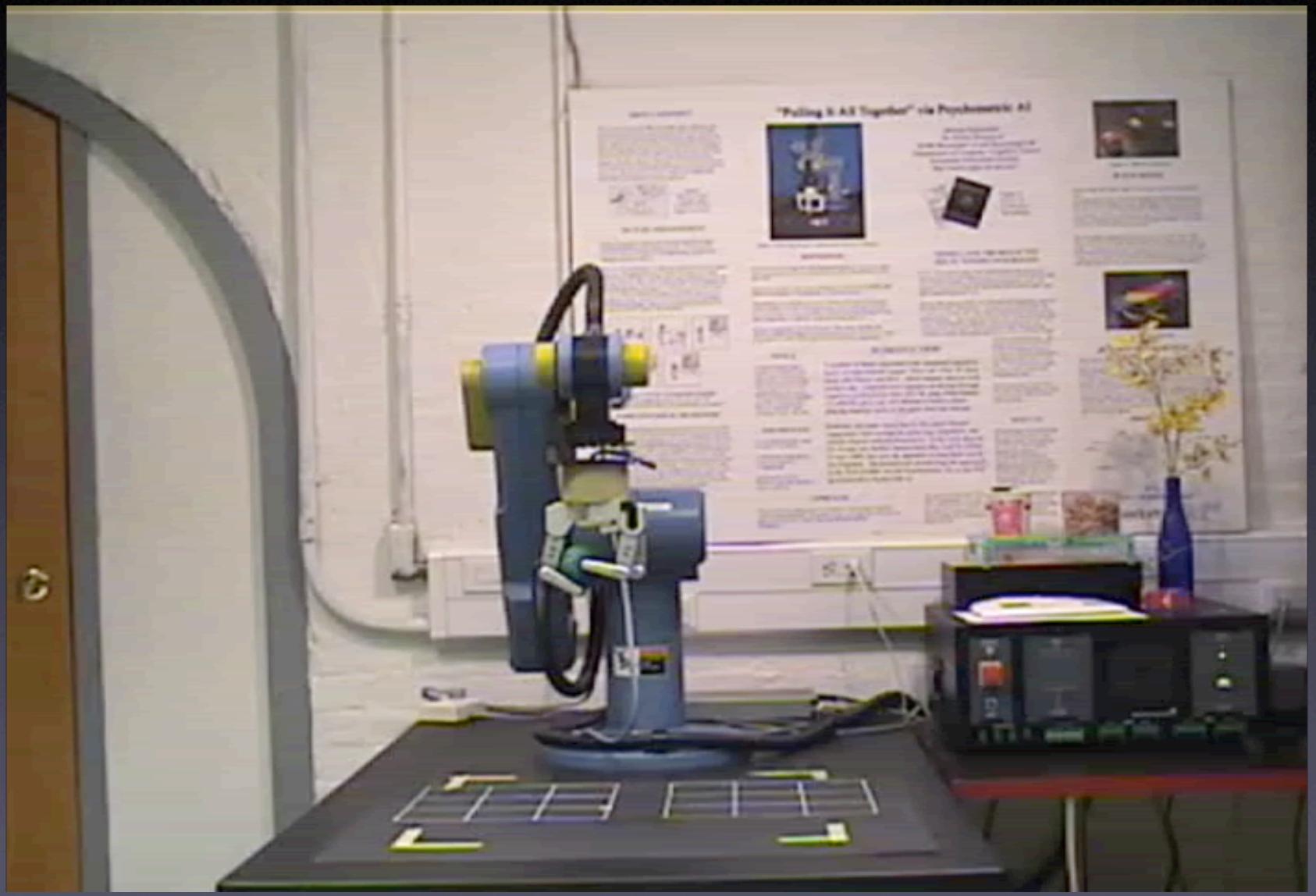
$$a_1(t+1) = A_1(a_1(t), s_2(t))$$
$$a_2(t+1) = A_2(a_2(t), s_1(t), s_3(t), s_{10}(t))$$
$$a_3(t+1) = A_3(a_3(t), s_4(t), s_5(t), s_6(t), s_8(t))$$
$$a_4(t+1) = A_4(a_4(t), s_7(t))$$



$$
\begin{aligned}
a_1(t+1) &= a_1(t) + s_2(t), \\
a_2(t+1) &= a_2(t) + s_1(t) + 2s_3(t), \\
a_3(t+1) &= \textbf{if } a_3(t) = 0 \textbf{ then } 0 \textbf{ else } a_3(t) + 1, \\
s_1(t) &= \textbf{if } a_1(t) = 0 \textbf{ then } 2 \textbf{ else } 1, \\
s_2(t) &= 1, \\
s_3(t) &= \textbf{if } a_3(t) = 0 \textbf{ then } 0 \textbf{ else } 1.
\end{aligned}
$$

(dirt-simple) *Theorem*. If in the initial state all subAutomata are in state 0, then subAutomaton 1 "can" put subAutomaton 2 in state 1, but won't. (I.e., if subAutomaton 1 emitted a 1 at time 0 instead of 2, subAutomaton 2 would go to state 1.)

Well, likewise, we have an existence proof that PERI can drop Earth, and can hold Earth, depending upon what automata-based functions produce as output, either `(drop-earth)`, or `(hold-earth)`.



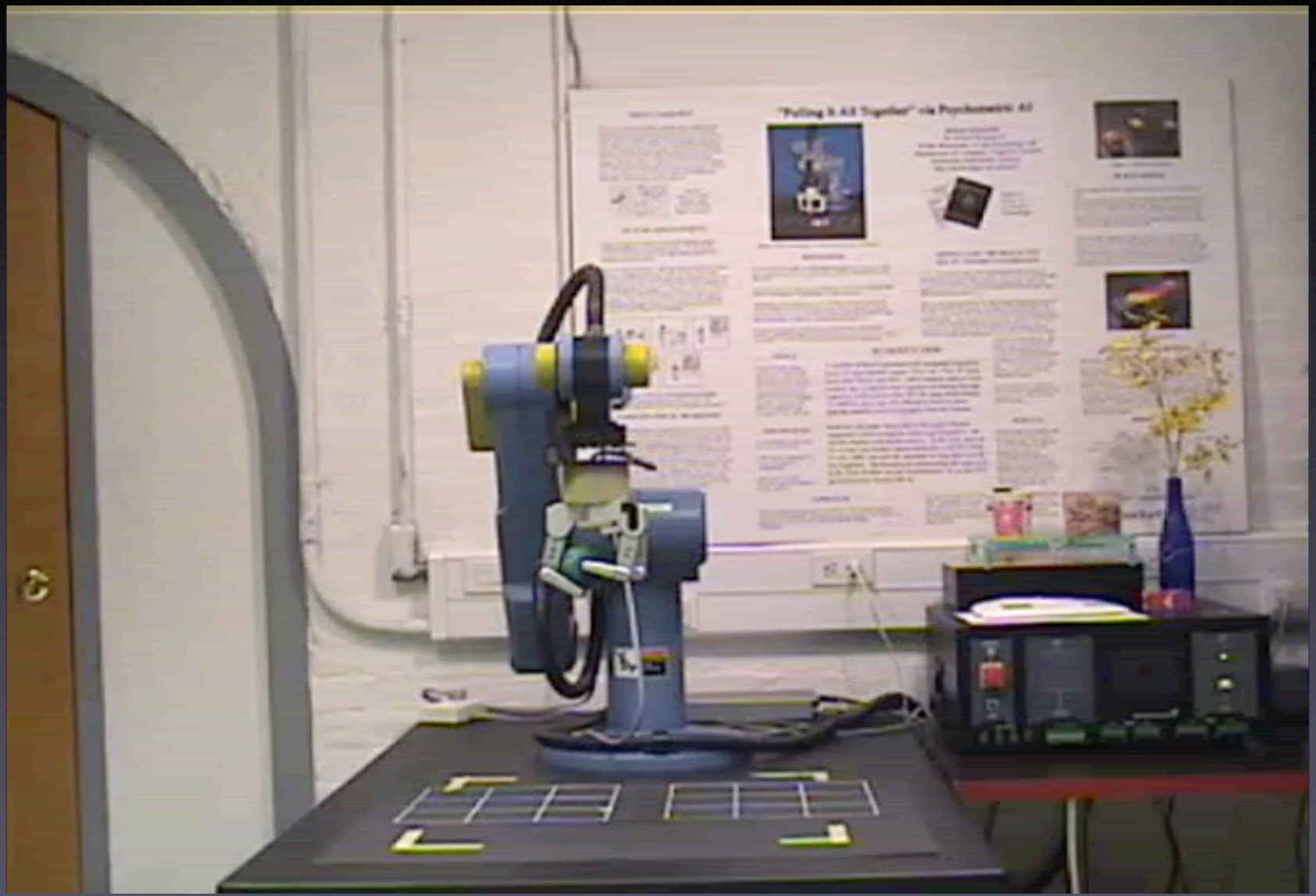
But this option fails to deliver, since, though PERI performs a morally permissible action, he doesn't *freely* perform it. In fact, the action isn't up to him at all, but is wholly dictated by a deterministic chain of events that *we* control.

Option #2 for PERI:

Decisions based on deduction over knowledge

$$Can(Result(a, s), s) \wedge [\forall s'(Can(s', s) \rightarrow s' <_{good} Result(a, s)] \rightarrow Should(a, s)$$

Once again we have an existence proof that PERI can drop Earth, and can hold Earth, depending upon what automata-based functions produce as output, either `(drop-earth)`, or `(hold-earth)`.

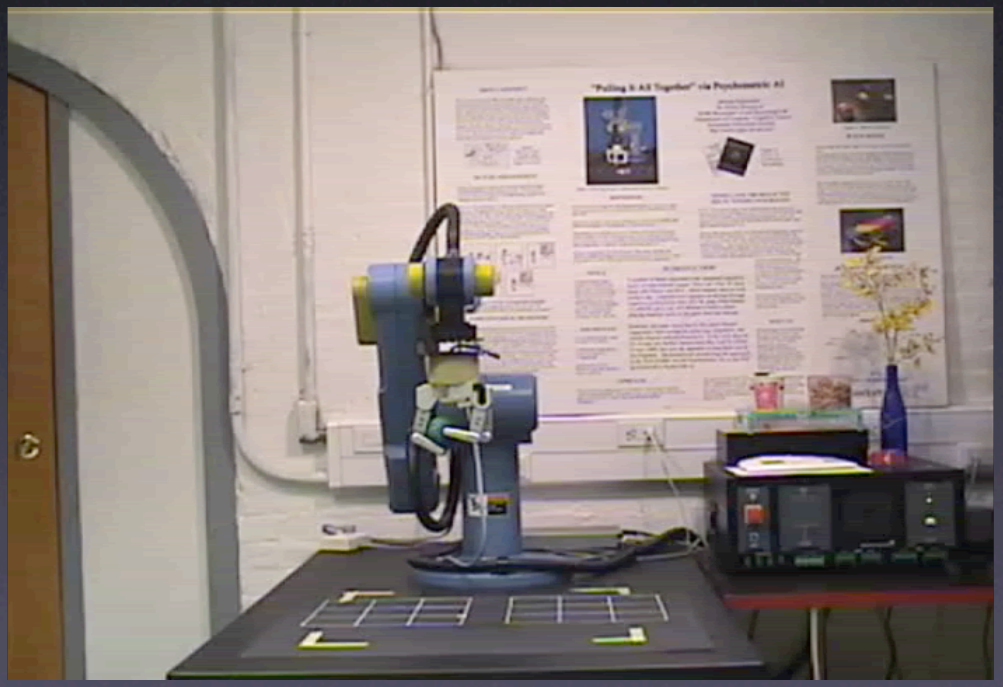


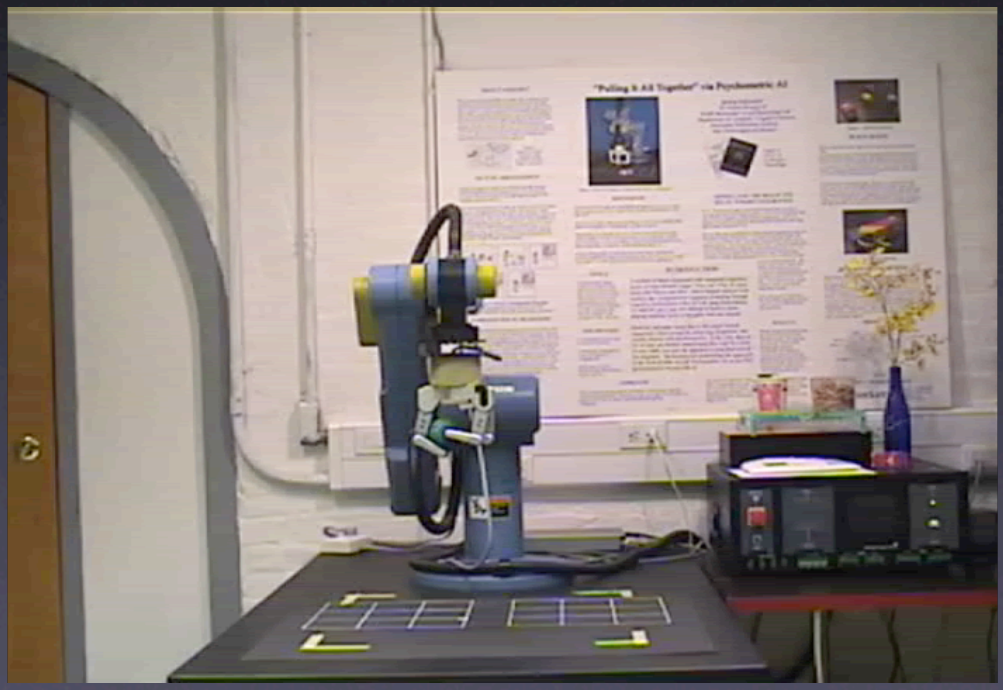
But this option also fails to deliver, since, once again, though PERI performs a morally permissible action, he doesn't *freely* perform it. In fact, the action isn't up to him at all, but is wholly dictated by a deterministic chain of events that *we* control.

Option #3: Exploit the random...

```
(defun peris-choice ()
       (cond ((> (random 10) 5) (hold-earth))
             ((drop-earth))))
```

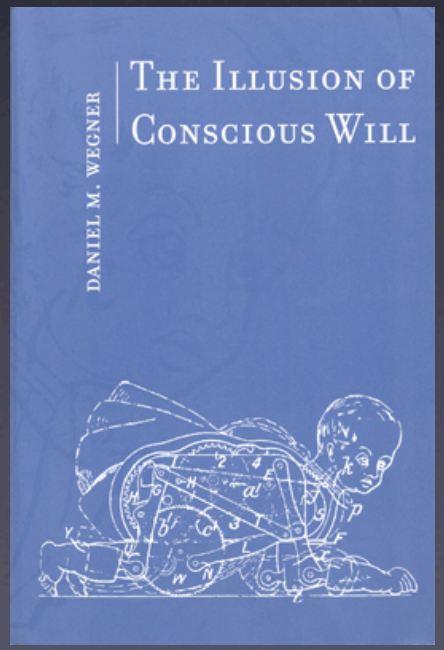
```
? (peris-choice)
"I will drop earth"

? (peris-choice)
"I will hold onto earth"

? (peris-choice)
"I will hold onto earth"
```



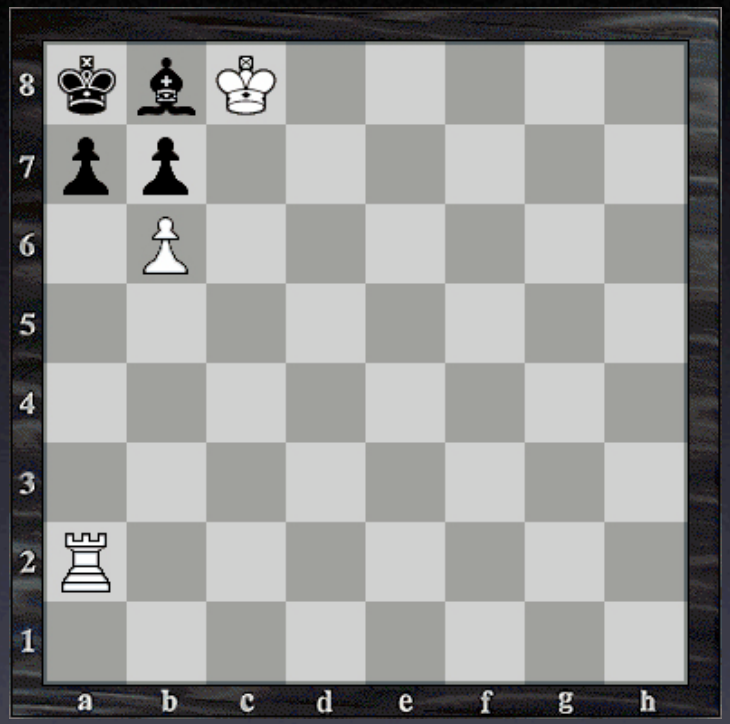


But this option *also* fails to deliver, since, once again, though PERI performs a morally permissible action, he doesn't *freely* perform it. In fact, the action isn't up to him at all, but is wholly capricious, dictated as it is by a process for generating random numbers outside of his control.

(This is by the way dreadfully bad news for those who infer from phenomena like hypnosis that the "concious will" is an illusion. It's a non sequitur to infer from what we have engineered under Option #3 that it's impossible for a robot to have freedom.)



Option #4: God as Cosmically Grand Grandmaster...



This is a chapter beyond today's limited scope.
But clearly, Option #4 is the next move for a Chisholmian "Agent Causation" man to explore.

The End

Analysis Needed For...

- *can, could've, ...* leave intuitive, for now
- *permissible, obligatory, forbidden* deontic logic; see AAAI *Machine Ethics* paper
- *freedom* now
  - in human persons
  - in robots

Desiderata for Deontic Logic-Controlled Ethical Robots

- All actions taken by robots are permissible.
- All obligatory actions for robots are performed by them (subject to ties and conflicts among available actions).
- No forbidden actions are performed by robots. (logically redundant).
- All permissible (or obligatory or forbidden) actions can be *proved* by the robot and in some cases, associated systems, e.g., oversight systems) to be permissible (or obligatory or forbidden), and all such proofs can be explained in ordinary English.