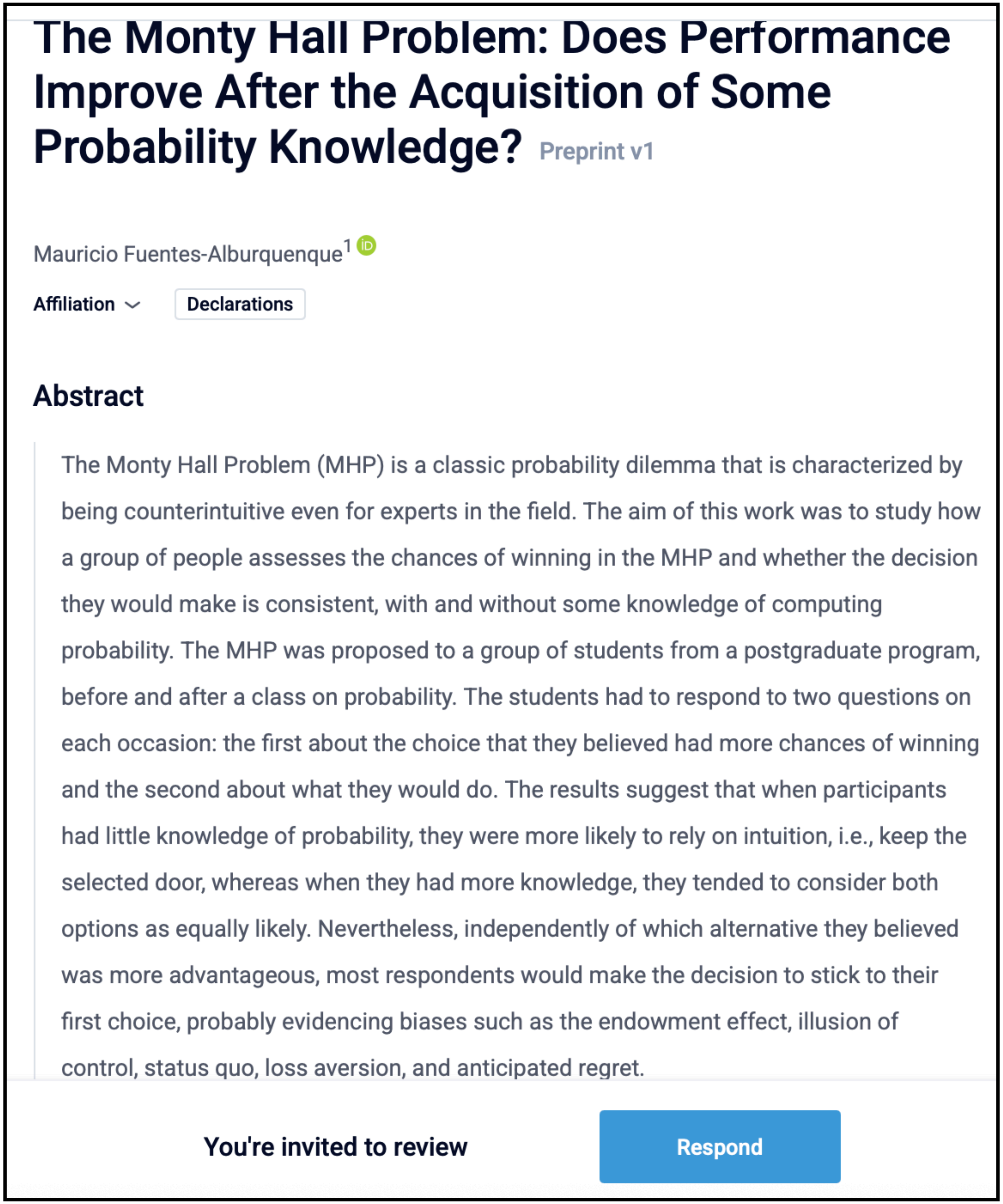
Computational-Logic Puzzlers Q&A

ILBAI

Selmer Bringsjord

11.20.25

Monty Hall Problems …



The Monty Hall Problem: Does Performance Improve After the Acquisition of Some Probability Knowledge? Preprint v1

Mauricio Fuentes-Alburquenque[1]

Abstract

The Monty Hall Problem (MHP) is a classic probability dilemma that is characterized by being counterintuitive even for experts in the field. The aim of this work was to study how a group of people assesses the chances of winning in the MHP and whether the decision they would make is consistent, with and without some knowledge of computing probability. The MHP was proposed to a group of students from a postgraduate program, before and after a class on probability. The students had to respond to two questions on each occasion: the first about the choice that they believed had more chances of winning and the second about what they would do. The results suggest that when participants had little knowledge of probability, they were more likely to rely on intuition, i.e., keep the selected door, whereas when they had more knowledge, they tended to consider both options as equally likely. Nevertheless, independently of which alternative they believed was more advantageous, most respondents would make the decision to stick to their first choice, probably evidencing biases such as the endowment effect, illusion of control, status quo, loss aversion, and anticipated regret.

Those who fail are behaving irrationally:

Friedman, D. (1998) “Monty Hall’s Three Doors: Construction and Deconstruction of a Choice Anomaly” *American Economic Review* **88**(4): 933–946.

- http://static.luiss.it/hey/ambiguity/papers/Friedman_1998.pdf

Those who fail are behaving irrationally:

Friedman, D. (1998) “Monty Hall’s Three Doors: Construction and Deconstruction of a Choice Anomaly” *American Economic Review* **88**(4): 933–946.

- http://static.luiss.it/hey/ambiguity/papers/Friedman_1998.pdf

Sometimes people make decisions that seem inconsistent with rational choice theory. We have a "choice anomaly" when such decisions are systematic and well documented. From a few isolated examples such as the Maurice Allais (1953) paradox and the probability matching puzzle of William K. Estes (1954), the set of anomalies expanded dramatically in the last two decades, especially following the work of Daniel Kahneman and Amos Tversky (e.g., 1979). By now the empirical literature offers dozens of interrelated anomalies documented in hundreds of articles and surveys (e.g., Colin F. Camerer, 1995).

Those who fail are behaving irrationally:

Friedman, D. (1998) “Monty Hall’s Three Doors: Construction and Deconstruction of a Choice Anomaly” *American Economic Review* **88**(4): 933–946.

- http://static.luiss.it/hey/ambiguity/papers/Friedman_1998.pdf

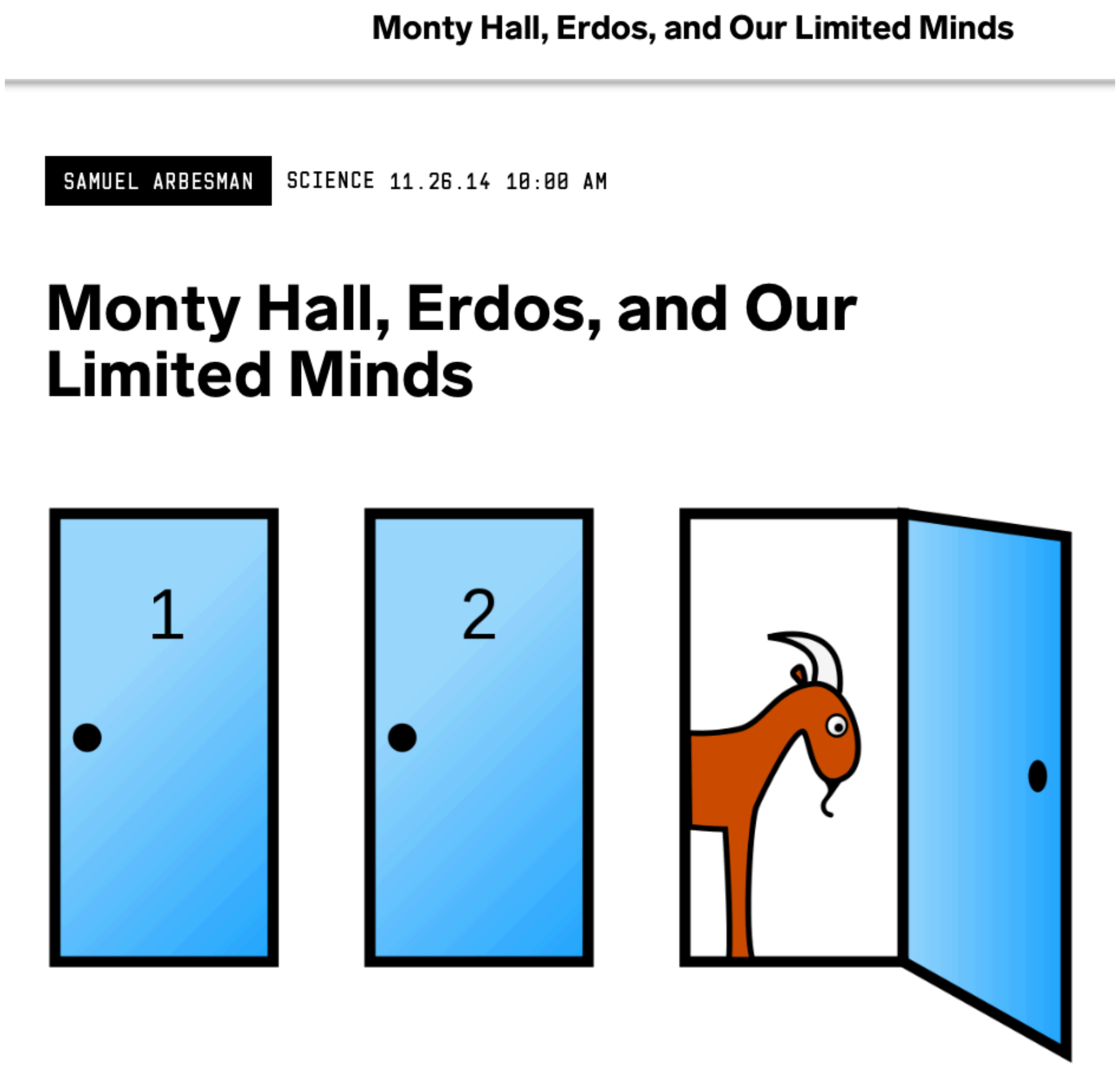
> Sometimes people make decisions that seem inconsistent with rational choice theory. We have a "choice anomaly" when such decisions are systematic and well documented. From a few isolated examples such as the Maurice Allais (1953) paradox and the probability matching puzzle of William K. Estes (1954), the set of anomalies expanded dramatically in the last two decades, especially following the work of Daniel Kahneman and Amos Tversky (e.g., 1979). By now the empirical literature offers dozens of interrelated anomalies documented in hundreds of articles and surveys (e.g., Colin F. Camerer, 1995).

Anomalies?? You mean irrational decisions?

Don’t Trust the Popular Media!

https://www.wired.com/2014/11/monty-hall-erdos-limited-minds/

Don’t Trust the Popular Media!



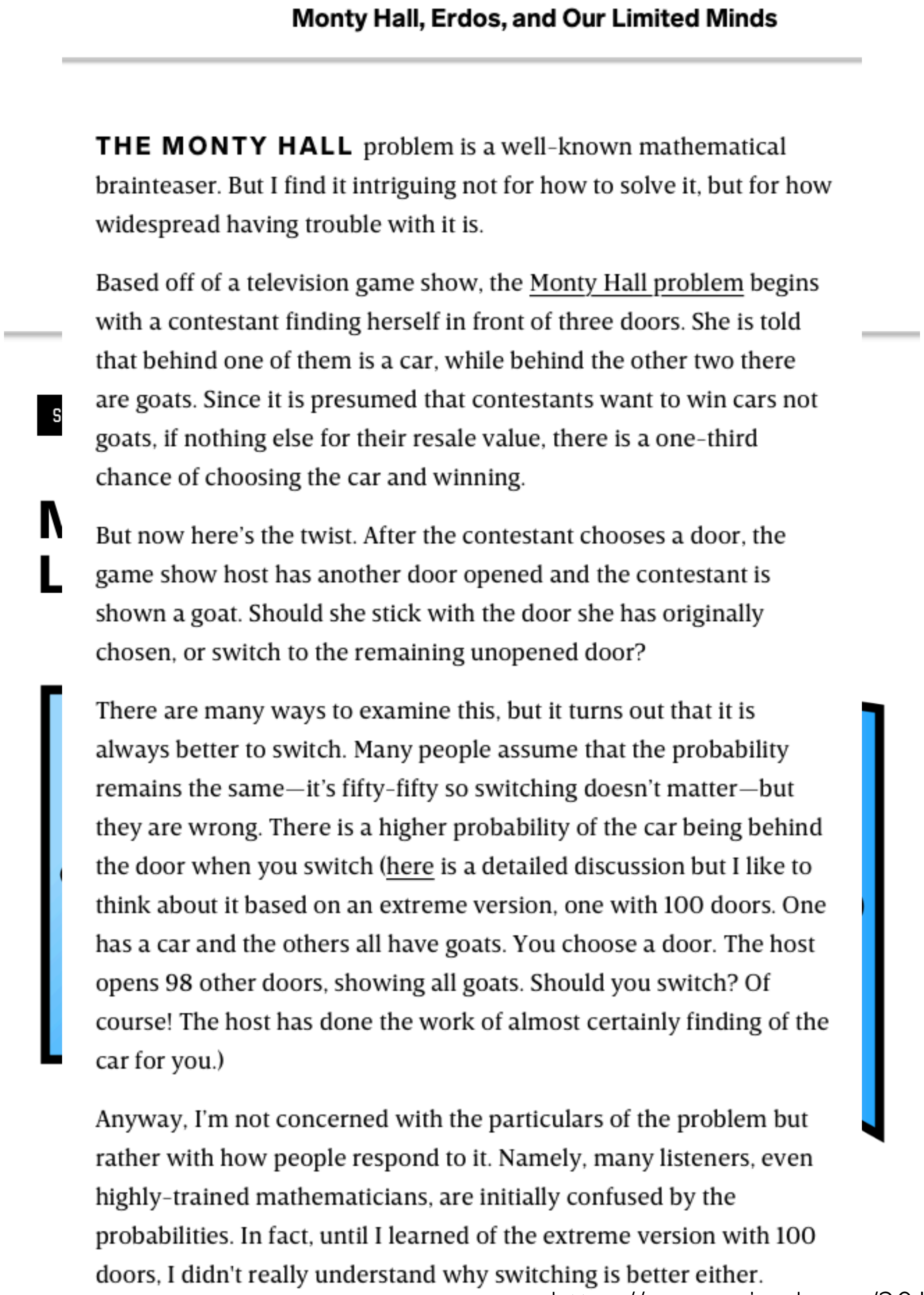
Monty Hall, Erdos, and Our Limited Minds

SAMUEL ARBESMAN SCIENCE 11.26.14 10:00 AM

Monty Hall, Erdos, and Our Limited Minds

https://www.wired.com/2014/11/monty-hall-erdos-limited-minds/

Don’t Trust the Popular Media!



Monty Hall, Erdos, and Our Limited Minds

THE MONTY HALL problem is a well-known mathematical brainteaser. But I find it intriguing not for how to solve it, but for how widespread having trouble with it is.

Based off of a television game show, the Monty Hall problem begins with a contestant finding herself in front of three doors. She is told that behind one of them is a car, while behind the other two there are goats. Since it is presumed that contestants want to win cars not goats, if nothing else for their resale value, there is a one-third chance of choosing the car and winning.

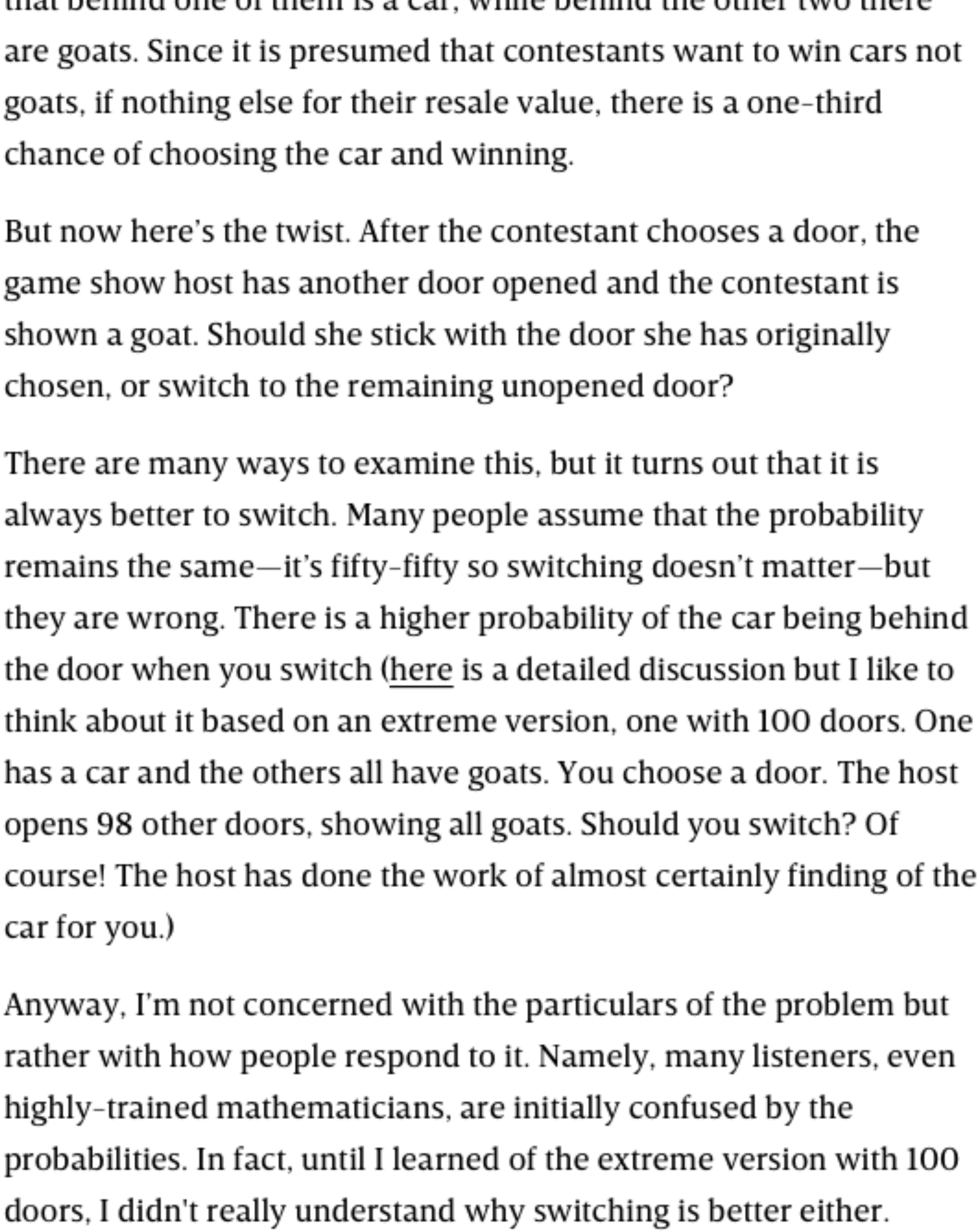
But now here’s the twist. After the contestant chooses a door, the game show host has another door opened and the contestant is shown a goat. Should she stick with the door she has originally chosen, or switch to the remaining unopened door?

There are many ways to examine this, but it turns out that it is always better to switch. Many people assume that the probability remains the same—it’s fifty-fifty so switching doesn’t matter—but they are wrong. There is a higher probability of the car being behind the door when you switch (here is a detailed discussion but I like to think about it based on an extreme version, one with 100 doors. One has a car and the others all have goats. You choose a door. The host opens 98 other doors, showing all goats. Should you switch? Of course! The host has done the work of almost certainly finding of the car for you.)

Anyway, I’m not concerned with the particulars of the problem but rather with how people respond to it. Namely, many listeners, even highly-trained mathematicians, are initially confused by the probabilities. In fact, until I learned of the extreme version with 100 doors, I didn't really understand why switching is better either.

https://www.wired.com/2014/11/monty-hall-erdos-limited-minds/

Don’ edia!



that behind one of them is a car, while behind the other two there are goats. Since it is presumed that contestants want to win cars not goats, if nothing else for their resale value, there is a one-third chance of choosing the car and winning.

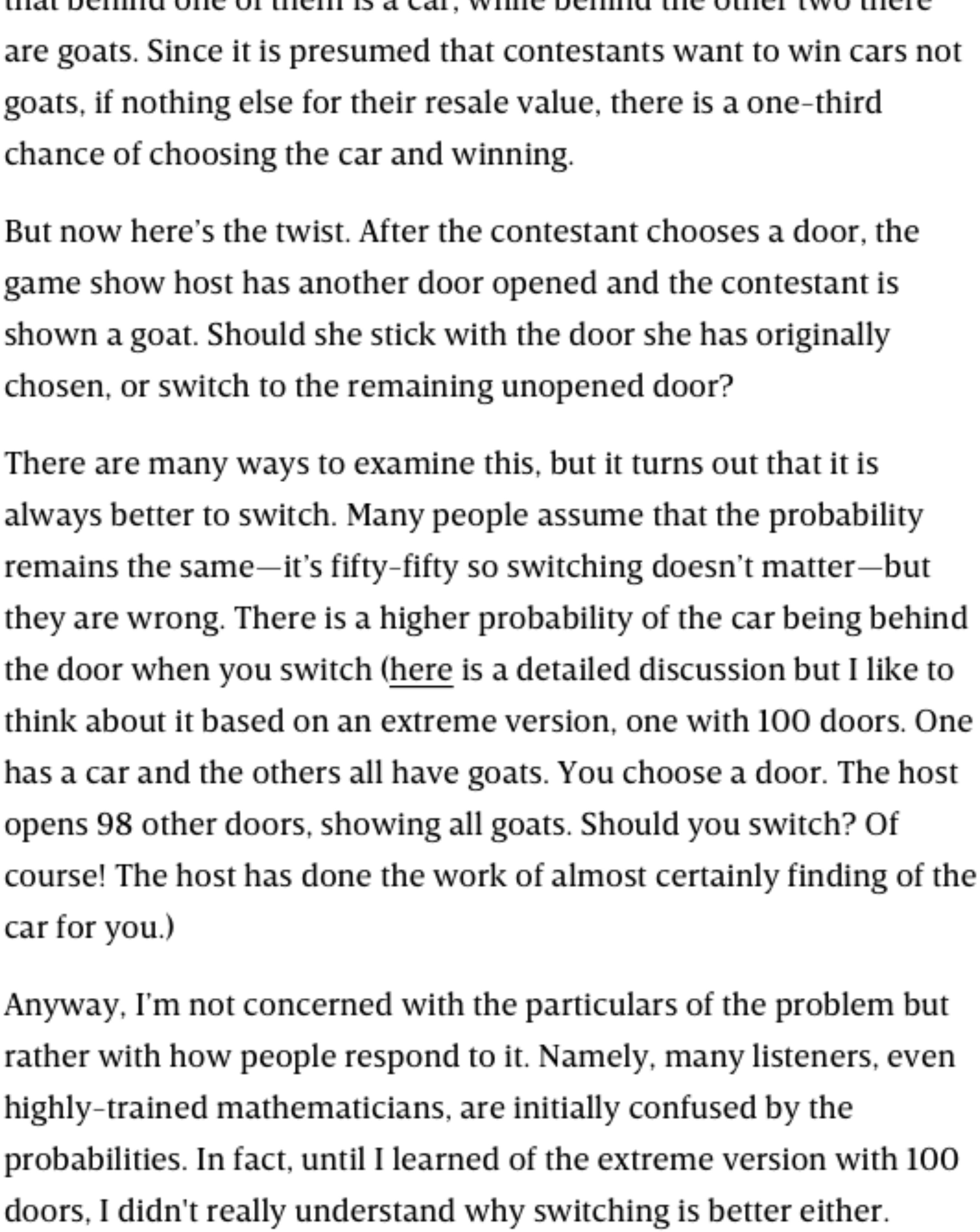
But now here’s the twist. After the contestant chooses a door, the game show host has another door opened and the contestant is shown a goat. Should she stick with the door she has originally chosen, or switch to the remaining unopened door?

There are many ways to examine this, but it turns out that it is always better to switch. Many people assume that the probability remains the same—it’s fifty-fifty so switching doesn’t matter—but they are wrong. There is a higher probability of the car being behind the door when you switch (here is a detailed discussion but I like to think about it based on an extreme version, one with 100 doors. One has a car and the others all have goats. You choose a door. The host opens 98 other doors, showing all goats. Should you switch? Of course! The host has done the work of almost certainly finding of the car for you.)

Anyway, I’m not concerned with the particulars of the problem but rather with how people respond to it. Namely, many listeners, even highly-trained mathematicians, are initially confused by the probabilities. In fact, until I learned of the extreme version with 100 doors, I didn't really understand why switching is better either.

https://www.wired.com/2014/11/monty-hall-erdos-limited-minds/

Don’ edia!



that behind one of them is a car, while behind the other two there are goats. Since it is presumed that contestants want to win cars not goats, if nothing else for their resale value, there is a one-third chance of choosing the car and winning.

But now here’s the twist. After the contestant chooses a door, the game show host has another door opened and the contestant is shown a goat. Should she stick with the door she has originally chosen, or switch to the remaining unopened door?

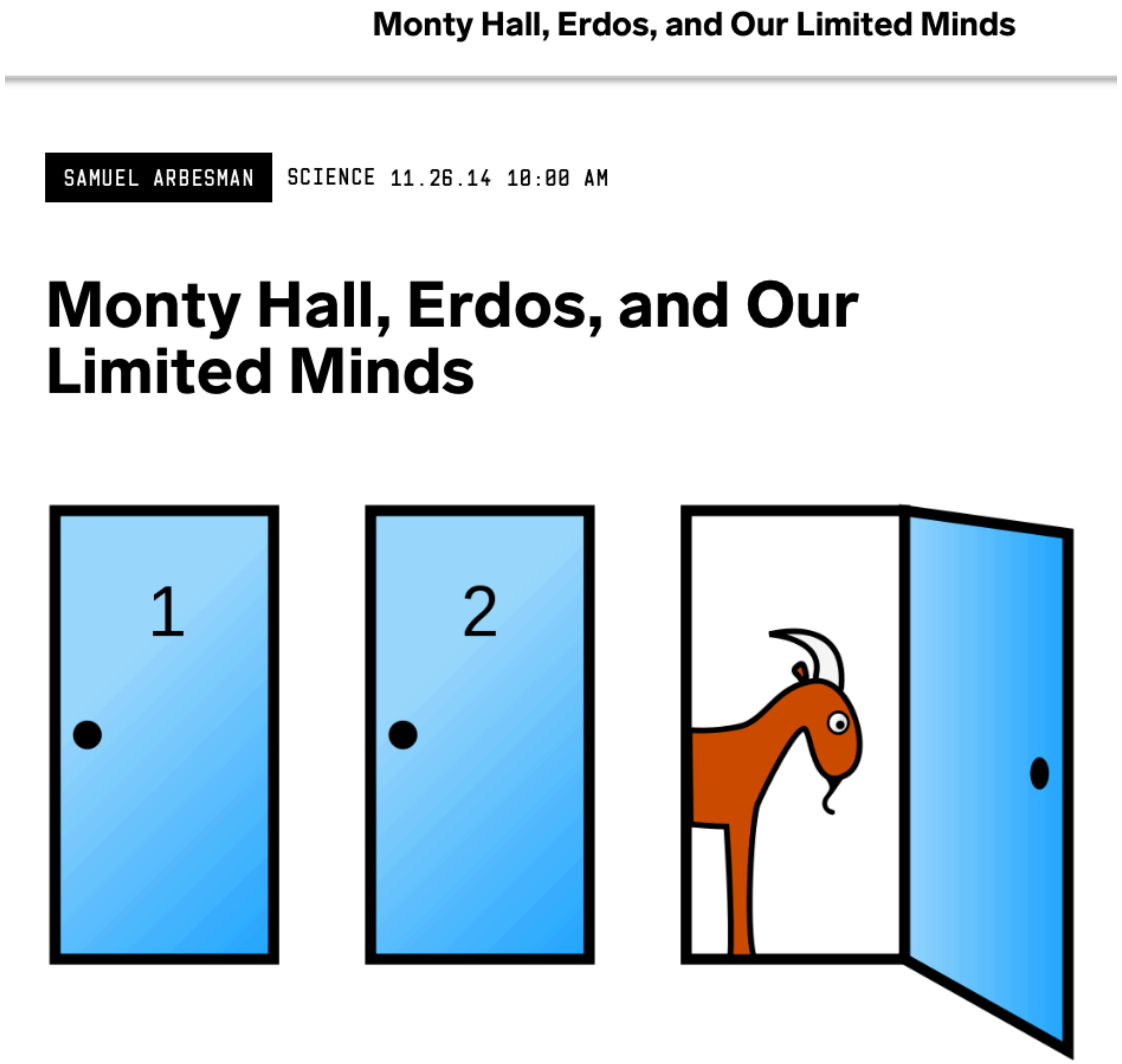
There are many ways to examine this, but it turns out that it is always better to switch. Many people assume that the probability remains the same—it’s fifty-fifty so switching doesn’t matter—but they are wrong. There is a higher probability of the car being behind the door when you switch (here is a detailed discussion but I like to think about it based on an extreme version, one with 100 doors. One has a car and the others all have goats. You choose a door. The host opens 98 other doors, showing all goats. Should you switch? Of course! The host has done the work of almost certainly finding of the car for you.)

Anyway, I’m not concerned with the particulars of the problem but rather with how people respond to it. Namely, many listeners, even highly-trained mathematicians, are initially confused by the probabilities. In fact, until I learned of the extreme version with 100 doors, I didn't really understand why switching is better either.

Painful!

https://www.wired.com/2014/11/monty-hall-erdos-limited-minds/

Don’t Trust Lazy Mathematicians!



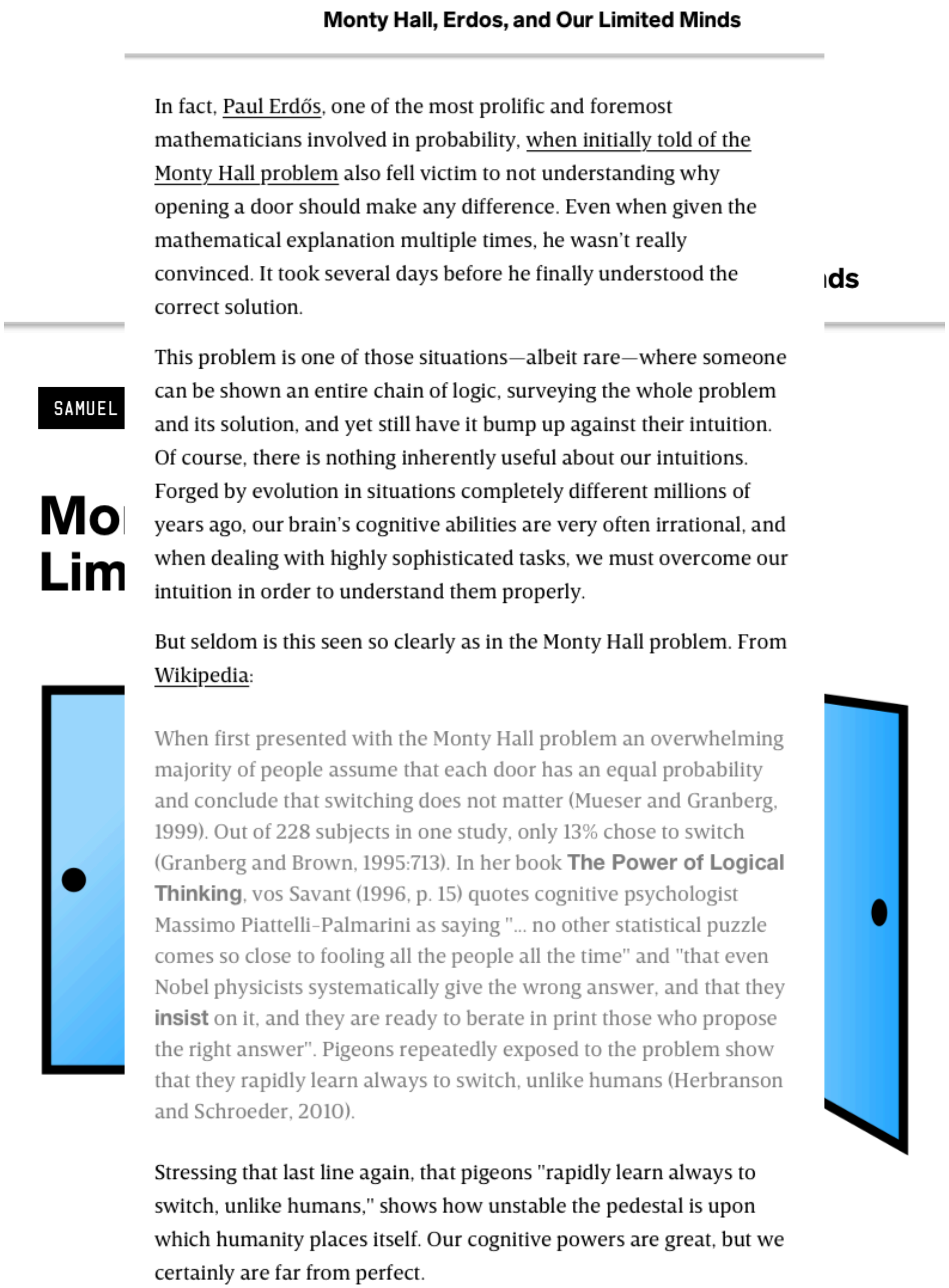
Monty Hall, Erdos, and Our Limited Minds

SAMUEL ARBESMAN SCIENCE 11.26.14 10:00 AM

Monty Hall, Erdos, and Our Limited Minds

https://www.wired.com/2014/11/monty-hall-erdos-limited-minds/

Don’t Trust Lazy Mathematicians!



Monty Hall, Erdos, and Our Limited Minds

In fact, Paul Erdős, one of the most prolific and foremost mathematicians involved in probability, when initially told of the Monty Hall problem also fell victim to not understanding why opening a door should make any difference. Even when given the mathematical explanation multiple times, he wasn’t really convinced. It took several days before he finally understood the correct solution.

This problem is one of those situations—albeit rare—where someone can be shown an entire chain of logic, surveying the whole problem and its solution, and yet still have it bump up against their intuition. Of course, there is nothing inherently useful about our intuitions. Forged by evolution in situations completely different millions of years ago, our brain’s cognitive abilities are very often irrational, and when dealing with highly sophisticated tasks, we must overcome our intuition in order to understand them properly.

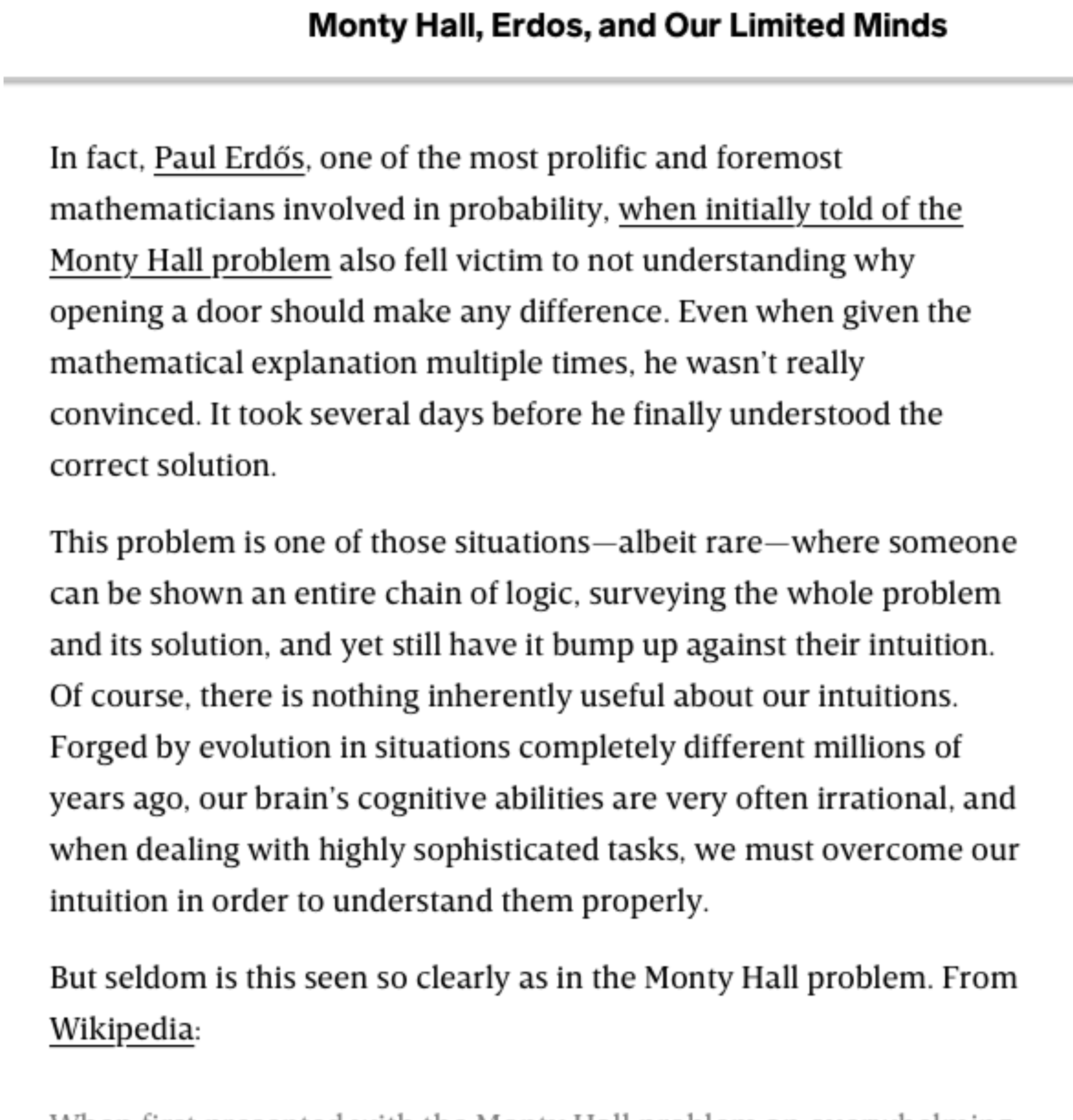
But seldom is this seen so clearly as in the Monty Hall problem. From Wikipedia:

> When first presented with the Monty Hall problem an overwhelming majority of people assume that each door has an equal probability and conclude that switching does not matter (Mueser and Granberg, 1999). Out of 228 subjects in one study, only 13% chose to switch (Granberg and Brown, 1995:713). In her book **The Power of Logical Thinking**, vos Savant (1996, p. 15) quotes cognitive psychologist Massimo Piattelli-Palmarini as saying "... no other statistical puzzle comes so close to fooling all the people all the time" and "that even Nobel physicists systematically give the wrong answer, and that they **insist** on it, and they are ready to berate in print those who propose the right answer". Pigeons repeatedly exposed to the problem show that they rapidly learn always to switch, unlike humans (Herbranson and Schroeder, 2010).

Stressing that last line again, that pigeons "rapidly learn always to switch, unlike humans," shows how unstable the pedestal is upon which humanity places itself. Our cognitive powers are great, but we certainly are far from perfect.

https://www.wired.com/2014/11/monty-hall-erdos-limited-minds/

Don’t Trust Lazy Mathematicians!



Monty Hall, Erdos, and Our Limited Minds

In fact, Paul Erdős, one of the most prolific and foremost mathematicians involved in probability, when initially told of the Monty Hall problem also fell victim to not understanding why opening a door should make any difference. Even when given the mathematical explanation multiple times, he wasn’t really convinced. It took several days before he finally understood the correct solution.

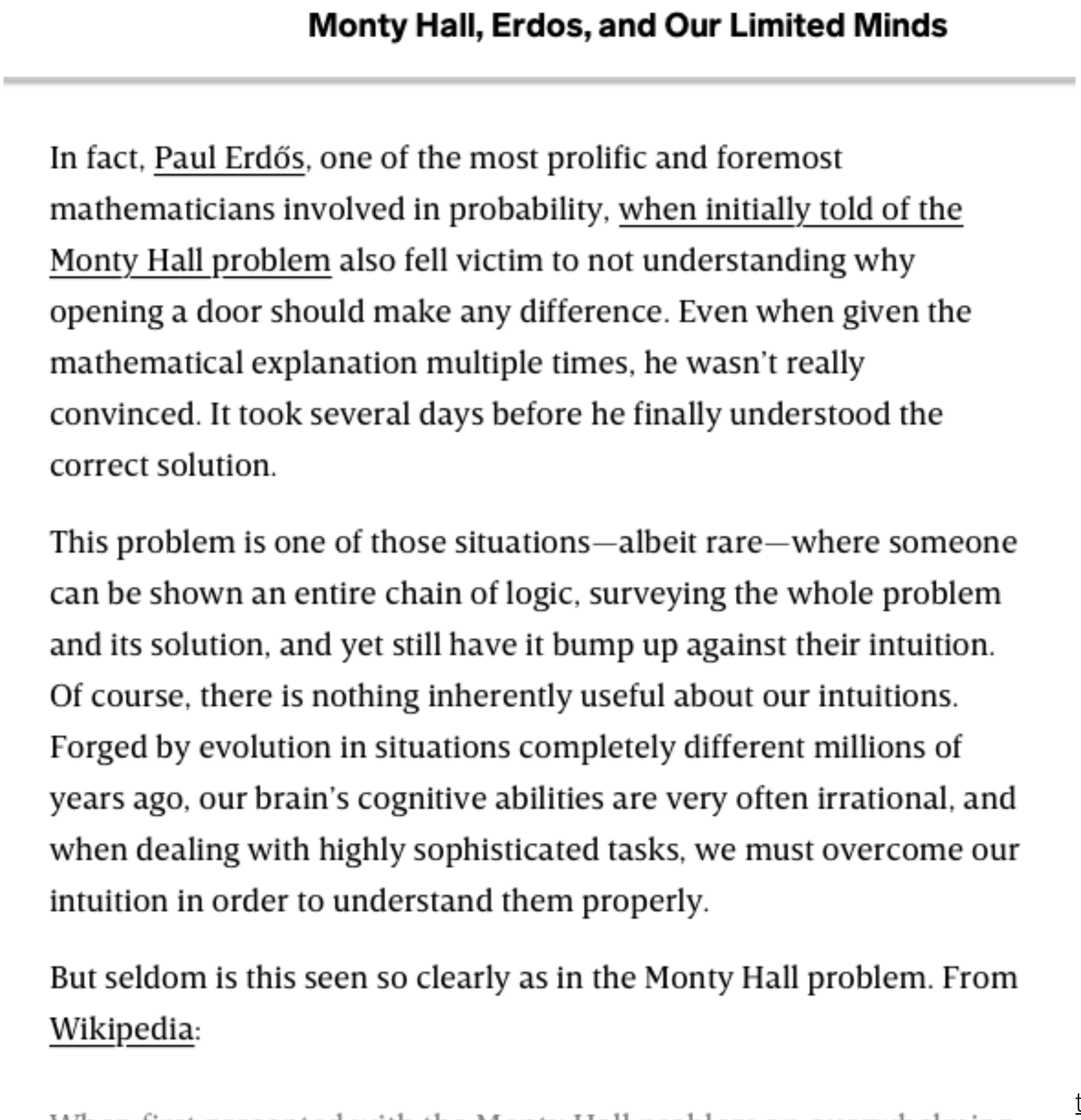
This problem is one of those situations—albeit rare—where someone can be shown an entire chain of logic, surveying the whole problem and its solution, and yet still have it bump up against their intuition. Of course, there is nothing inherently useful about our intuitions. Forged by evolution in situations completely different millions of years ago, our brain’s cognitive abilities are very often irrational, and when dealing with highly sophisticated tasks, we must overcome our intuition in order to understand them properly.

But seldom is this seen so clearly as in the Monty Hall problem. From Wikipedia:

When first presented with the Monty Hall problem an overwhelming

ty-hall-erdos-limited-minds/

Don’t Trust Lazy Mathematicians!



Monty Hall, Erdos, and Our Limited Minds

In fact, Paul Erdős, one of the most prolific and foremost mathematicians involved in probability, when initially told of the Monty Hall problem also fell victim to not understanding why opening a door should make any difference. Even when given the mathematical explanation multiple times, he wasn’t really convinced. It took several days before he finally understood the correct solution.

This problem is one of those situations—albeit rare—where someone can be shown an entire chain of logic, surveying the whole problem and its solution, and yet still have it bump up against their intuition. Of course, there is nothing inherently useful about our intuitions. Forged by evolution in situations completely different millions of years ago, our brain’s cognitive abilities are very often irrational, and when dealing with highly sophisticated tasks, we must overcome our intuition in order to understand them properly.

But seldom is this seen so clearly as in the Monty Hall problem. From Wikipedia:

When first presented with the Monty Hall problem an overwhelming

Gotta prove it, Paul!

ty-hall-erdos-limited-minds/

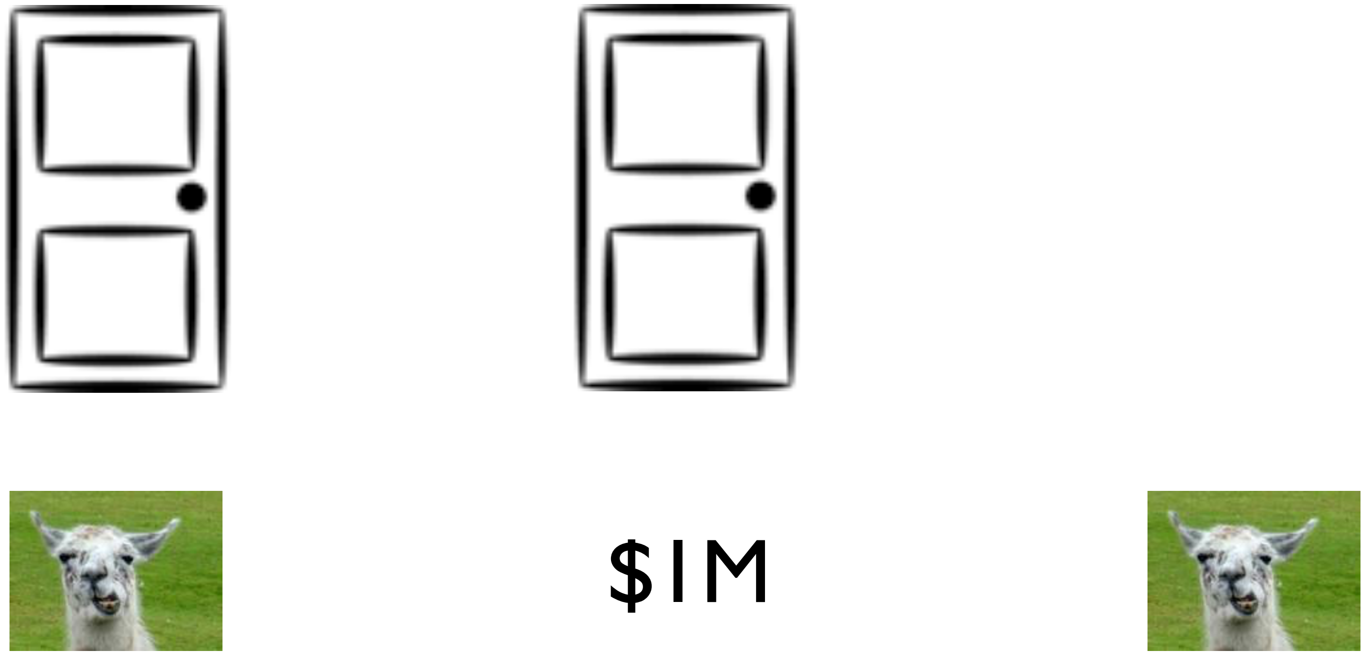
The Monty Hall Problem

$1M

The Monty Hall Problem

$1M

The Monty Hall Problem



$1M

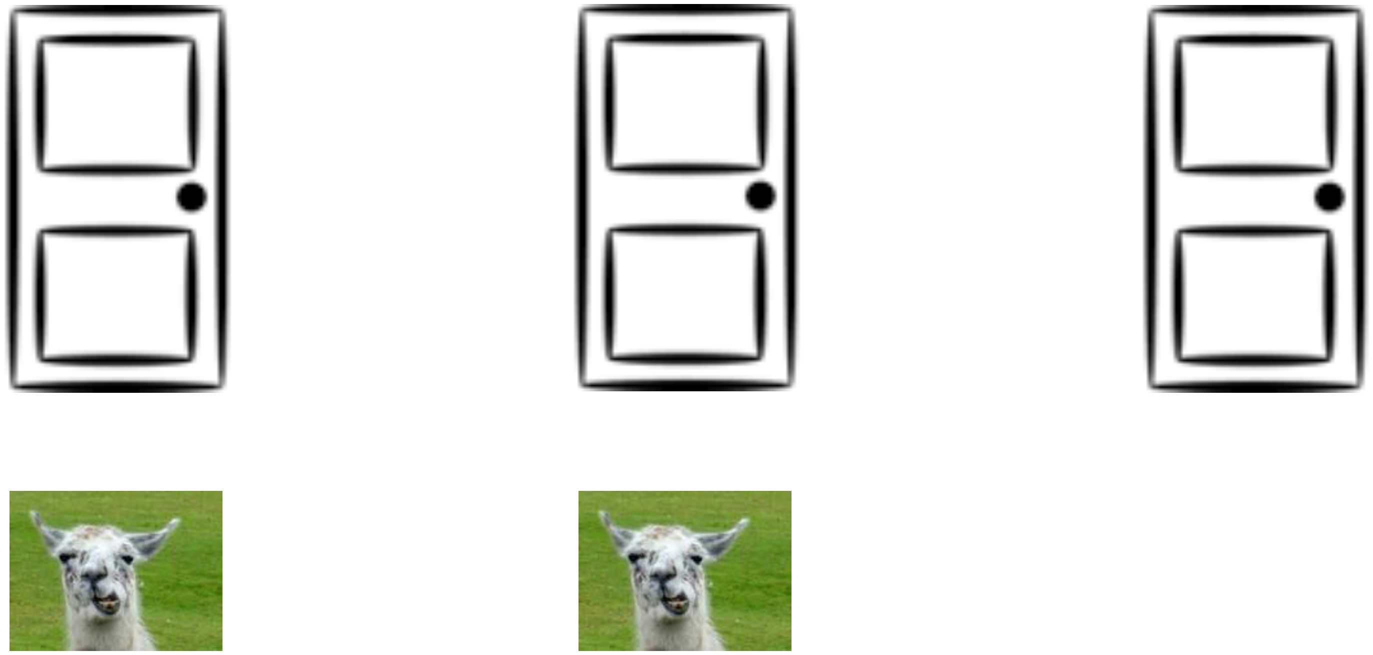
The Monty Hall Problem

$1M

The Monty Hall Problem

$1M

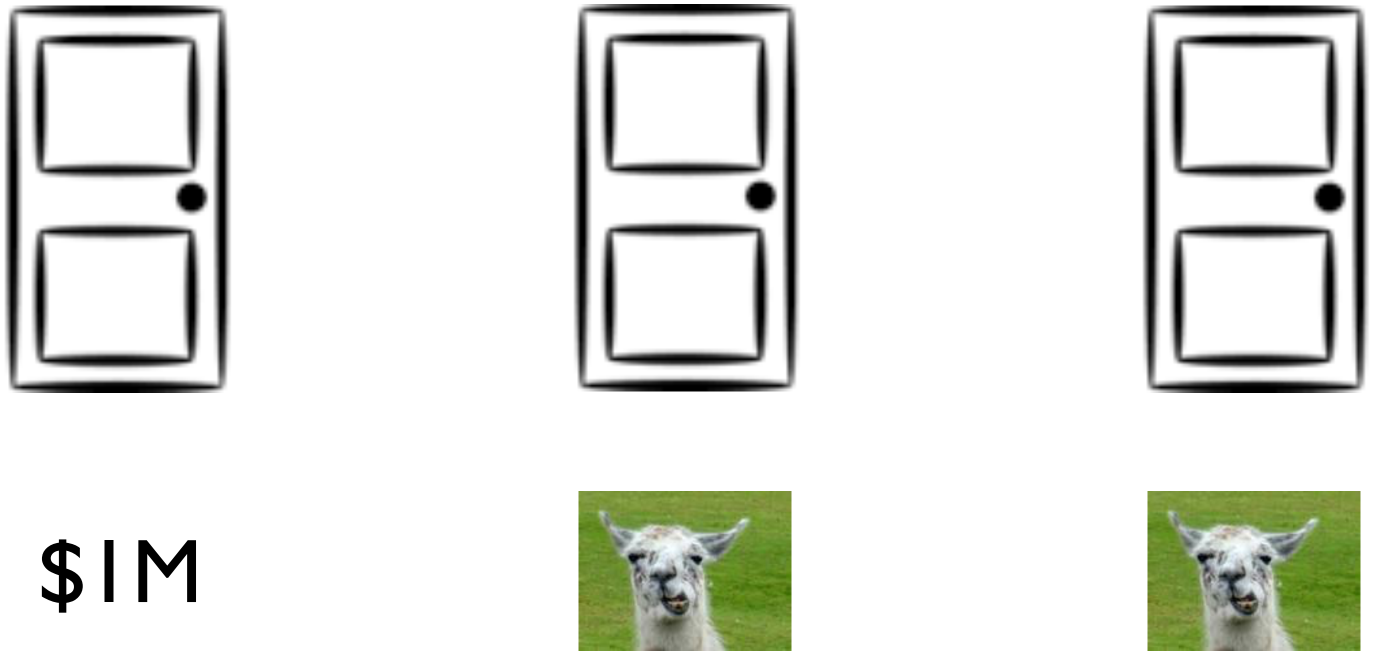
The Monty Hall Problem



The Monty Hall Problem

$1M

The Monty Hall Problem



MHP Defined

Jones has come to a game show, and finds himself thereon selected to play a game on national TV with the show's suave host, Full Monty. Jones is told correctly by Full that hidden behind one of three closed, opaque doors facing the two of them is $1,000,000, while behind each of the other two is a feculent, obstreperous llama whose value on the open market is charitably pegged at $1. Full reminds Jones that this is a game, and a fair one, and that if Jones ends up selecting the door with $1M behind it, all that money will indeed be his. (Jones' net worth has nearly been exhausted by his expenditures in traveling to the show.) Full also reminds Jones that he (= Full) knows what's behind each door, fixed in place until the game ends.

Full asks Jones to select which door he wants the contents of. Jones says, "Door 1." Full then says: "Hm. Okay. Part of this game is my revealing at this point what's behind one of the doors you didn't choose. So ... let me show you what's behind Door 3." Door 3 opens to reveal a very unsavory llama. Full now to Jones: "Do you want to switch to Door 2, or stay with Door 1? You'll get what's behind the door of your choice, and our game will end." Full looks briefly into the camera, directly.

(P1.1) What should Jones do if he's rational?

(P1.2) Prove that your answer is correct. (Diagrammatic proofs are allowed.)

(P1.3) A quantitative hedge fund manager with a PhD in finance from Harvard zipped this email off to Full before Jones made his decision re. switching or not: "Switching would be a royal waste of time (and time is money!). Jones hasn't a doggone clue what's behind Door 1 or Door 2, and it's obviously a 50/50 chance to win whether he stands firm or switches. So the chap shouldn't switch!" Is the fund manager right? Prove that your diagnosis is correct.

(P1.4) Can these answers and proofs be exclusively Bayesian in nature?

Any questions about how the game is played?

Any questions about how the game is played?

Okay, some questions for you now, then …

The Switching Policy Rational!

Proof: Our overarching technique will be proof by cases.

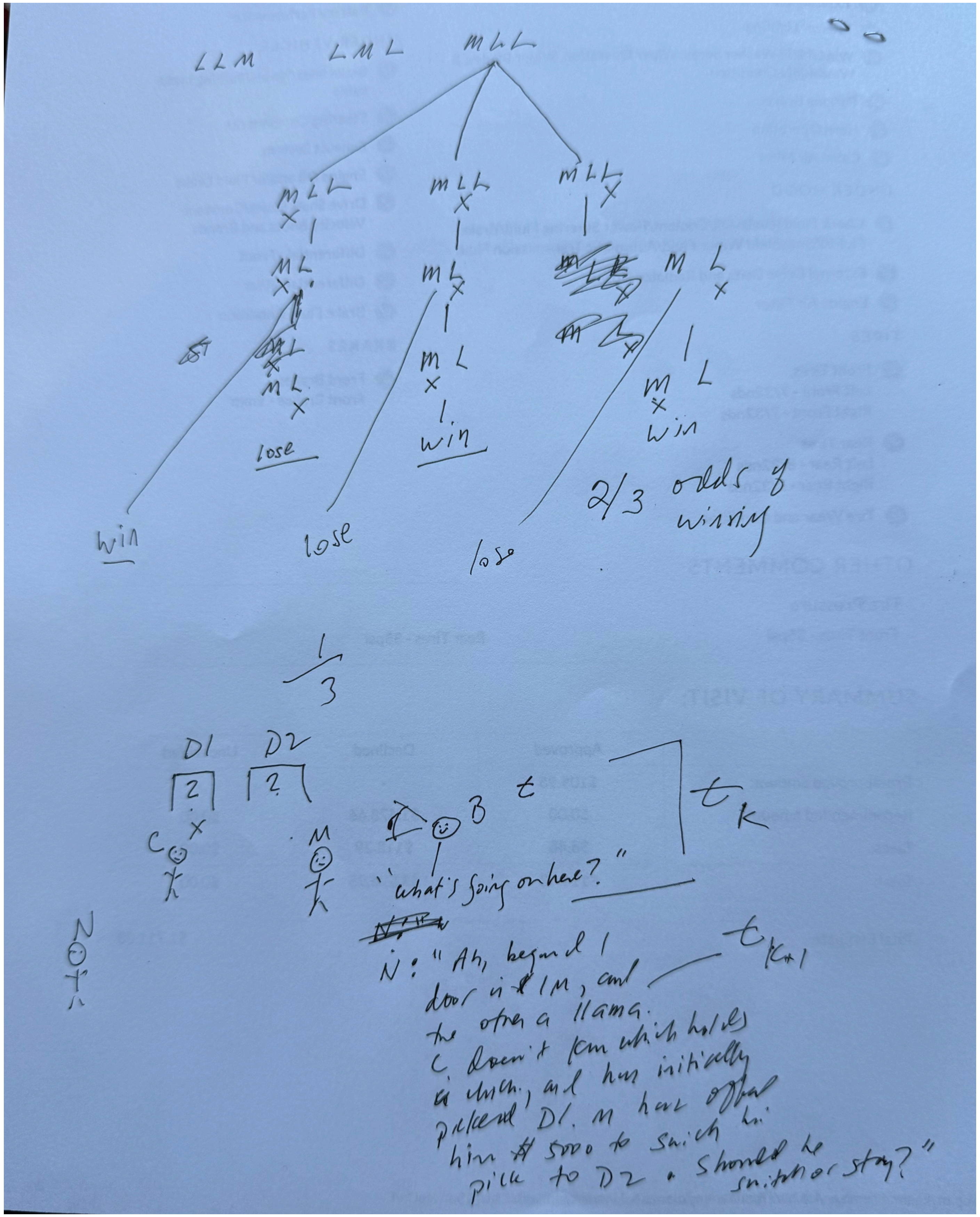
We denote the possible cases for initial distribution using a simple notation, according to which for example 'LLM' means that, there is a lama behind Door 1, a llama behind Door 2, and the million dollars behind Door 3. With this notation in hand, our three starting cases are: Case 1: MLL; Case 2: LML; Case 3: LLM. There are only three top-level cases for distribution. The odds of picking at the start the million-dollar door is 1/3, obviously — for each case. Hence we know that the odds of a HOLD policy winning is 1/3.

Now we proceed in a proof by sub-cases under the three cases above, to show that the overall odds of a SWITCH policy is greater than 1/3. Each sub-case is simply based on what the initial choice by Jones is, under one of the three main cases. Here we go:

Suppose Case 3, LLM, holds, and that [this (Case 3.1) is the first of three sub-cases under Case 3] Jones picks Door 1. Then FM must reveal Door 2 to reveal a llama. Switching to Door 3 wins, guaranteed. In sub-case 3.2 suppose that J's choice Door 2. Then FM will reveal Door 1. Again, switching to Door 3 wins, guaranteed. In the final sub-case, J initially selects Door 3 under Case 3; this is sub-case 3.3. Here, FM shows either Door 1 or Door 2 (as itself a random choice). This time switching loses, guaranteed. Hence, in two of the sub-cases out of three (2/3), winning is guaranteed (*prob* of 1). An exactly parallel result can be deduced for Case 2 and Case 1; i.e., in each of these two, in two of the three (2/3) sub-cases winning is 1. Hence the odds of winning by following the switching policy is 2/3, which is greater than 1/3. Hence it's rational to be a switcher. **QED**

- Walker-By 2-Door …

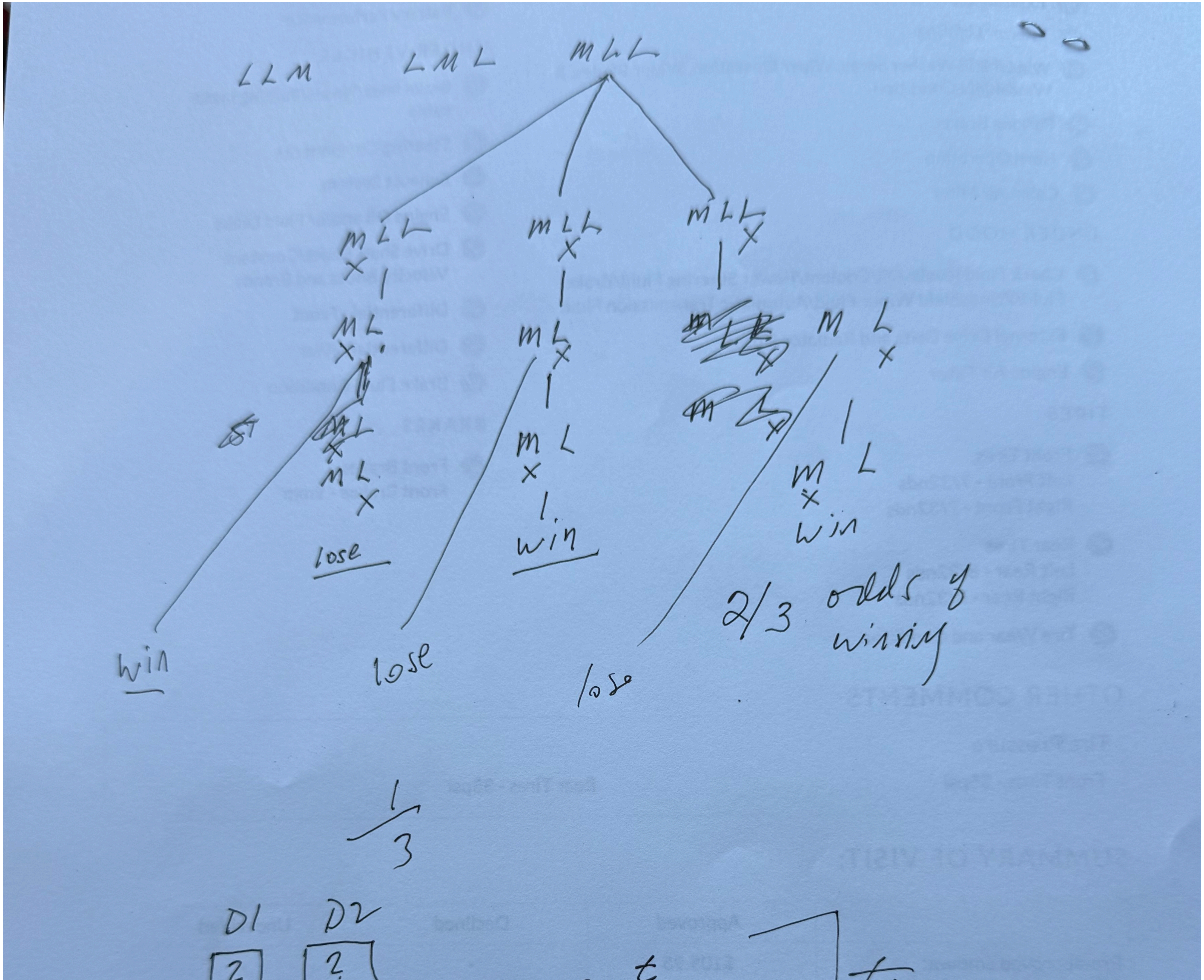
- Walker-By 2-Door …
- Random-Strike 3-Door …

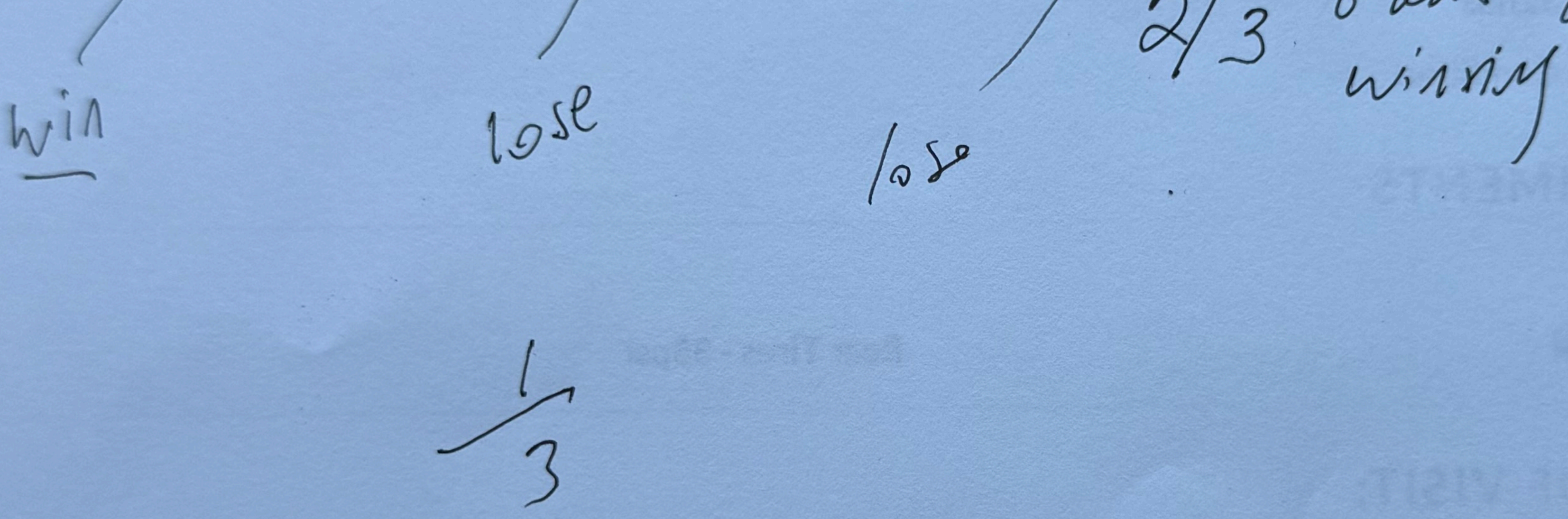


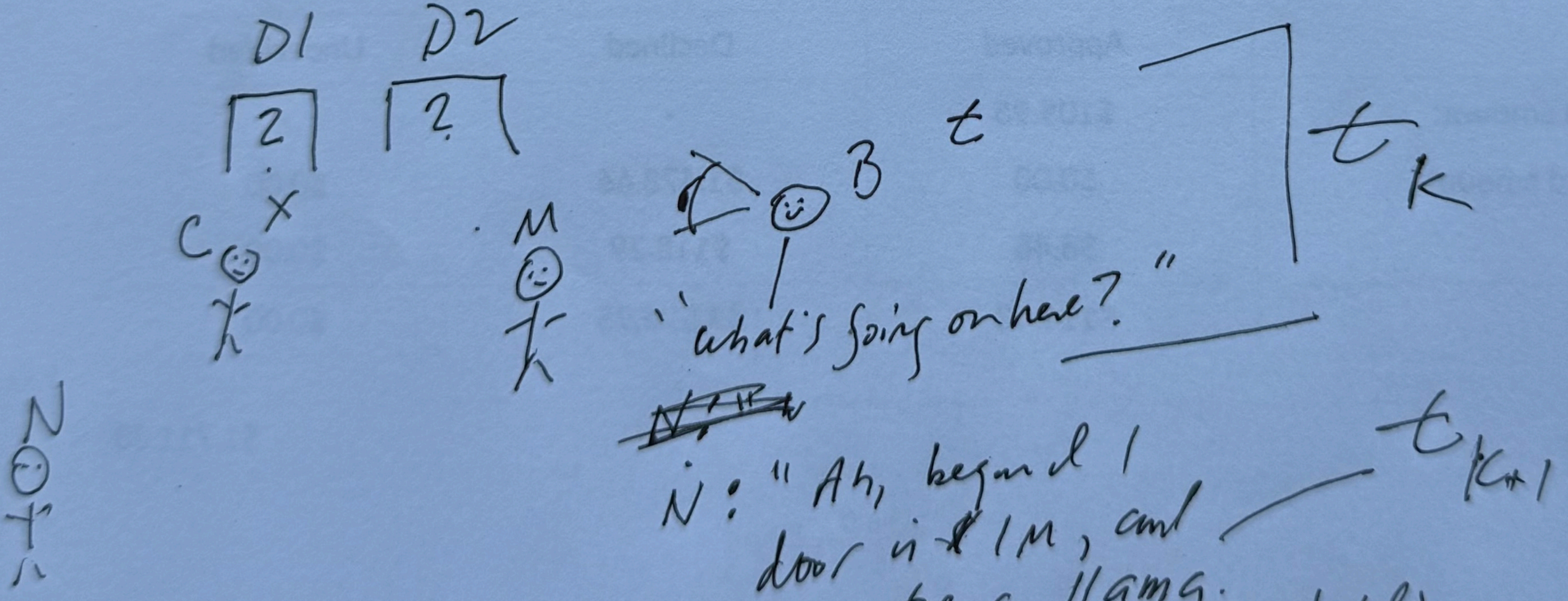
"what's going on here?"

~~N:~~

N: "Ah, beyond 1 door is 1M, and the other a llama. C doesn't know which holds which, and has initially picked D1. M has offered him $5000 to switch his pick to D2. Should he switch or stay?"


LLM
LML
MLL
MLL
×
ML
×
ML
×
lose
MLL
×
ML
×
ML
×
l
win
MLL
×
ML
×
L
M
×
win
win
lose
lose
2/3 odds of winning
1/3
D1
D2
?
?





~~N...~~

N: "Ah, beyond 1 door is $1M, and the other a llama. C doesn't know which holds ~~is~~ which, and has initially picked D1. M has offered him $5000 to switch his pick to D2. Should he switch or stay?"

What on Newcomb’s Problem …

(Review needed?)

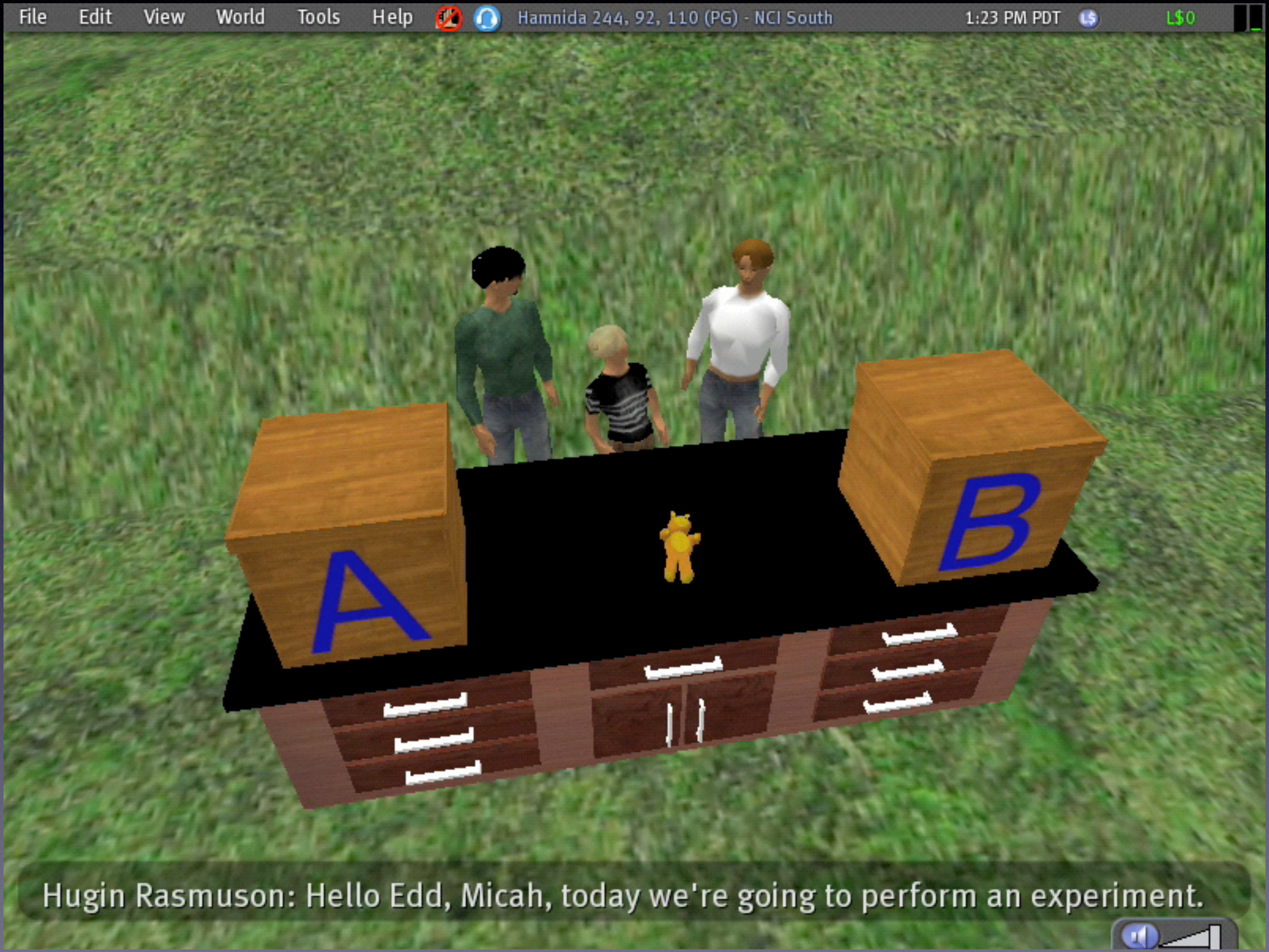
Floridi’s Continuum (augmented), and Claims

(“Consciousness, Agents, and the Knowledge Game” *Minds & Machines*)

	False Belief Task	Wise Man Test (*n*)	Deafening Test	Torture Boots Test	Ultimate Sifter	*Infinitary* False Belief Task
Cutting-Edge AI	Yes	Yes	No	No	No	?
Zombies	Yes	Yes	Yes	Yes	No	?
Human Persons (s-conscious! p-conscious!)	Yes	Yes	Yes	Yes	Yes	Yes

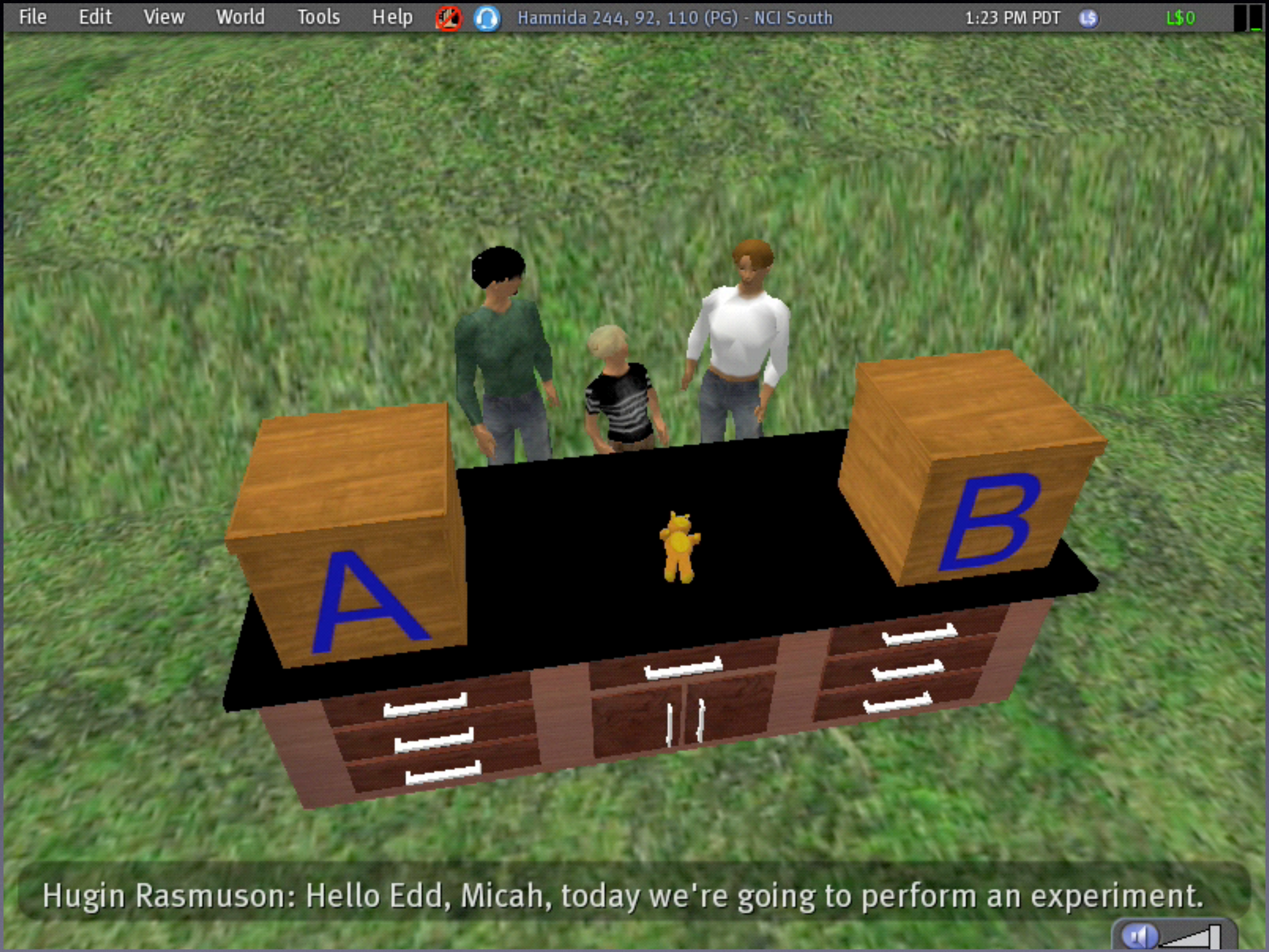
Cracking False-Belief Tasks ...

In *SL*, w/ real-time comm w/ ATP



Hugin Rasmuson: Hello Edd, Micah, today we're going to perform an experiment.

In *SL*, w/ real-time comm w/ ATP



Hugin Rasmuson: Hello Edd, Micah, today we're going to perform an experiment.

```
SNARK-USER 14 >
(in-immature-scenario
  (prove '(t-retrieve subject
                      teddybear
                      ?c)
         :answer '(looks-in ?c)))
```

```
(Refutation
(Row 1
   (or (not (person ?x)) (not (object ?y))
(not (container ?z)) (not (in ?y ?z))
(bel-in ?x ?y ?z))
   assertion)
(Row 2
   (or (not (person ?x))
       (not (container ?y))
       (not (object ?z))
       (not (w-retrieve ?x ?z))
       (not (bel-in ?x ?z ?y))
       (t-retrieve ?x ?z ?y))
   assertion)
(Row 4
   (person subject)
   assertion)
(Row 6
   (container c2)
   assertion)
(Row 7
   (object teddybear)
   assertion)
(Row 8
   (in teddybear c2)
   assertion)
(Row 9
   (w-retrieve subject teddybear)
   assertion)
(Row 10
   (not (t-retrieve subject teddybear ?x))
   negated_conjecture
   Answer (looks-in ?x))
(Row 11
   (or (not (person ?x)) (bel-in ?x
teddybear c2))
   (rewrite (resolve 1 8) 6 7))
(Row 25
   (bel-in subject teddybear c2)
   (resolve 11 4))
(Row 28
   (t-retrieve subject teddybear c2)
   (rewrite (resolve 2 25) 9 7 6 4))
(Row 30
   false
   (resolve 10 28)
   Answer (looks-in c2)))

:PROOF-FOUND

SNARK-USER 15 > (answer t)
(LOOKS-IN C2)
```

```
SNARK-USER 12 >
(in-mature-scenario
  (prove '(t-retrieve subject
                      teddybear
                      ?c)
         :answer '(looks-in ?c)))
```

```
(Refutation
(Row 1
   (or (not (person ?x))
       (not (container ?y))
       (not (object ?z))
       (not (w-retrieve ?x ?z))
       (not (bel-in ?x ?z ?y))
       (t-retrieve ?x ?z ?y))
   assertion)
(Row 2
   (or (not (person ?x)) (not (object ?
y)) (not (container ?z)) (not (p-in ?x ?y
?z)) (bel-in ?x ?y ?z))
   assertion)
(Row 4
   (person subject)
   assertion)
(Row 5
   (container c1)
   assertion)
(Row 7
   (object teddybear)
   assertion)
(Row 8
   (p-in subject teddybear c1)
   assertion)
(Row 9
   (w-retrieve subject teddybear)
   assertion)
(Row 10
   (not (t-retrieve subject teddybear ?
x))
   negated_conjecture
   Answer (looks-in ?x))
(Row 11
   (bel-in subject teddybear c1)
   (rewrite (resolve 2 8) 5 7 4))
(Row 25
   (t-retrieve subject teddybear c1)
   (rewrite (resolve 1 11) 9 7 5 4))
(Row 26
   false
   (resolve 10 25)
   Answer (looks-in c1))
)

:PROOF-FOUND

SNARK-USER 13 > (answer t)
(LOOKS-IN C1)
```

“The present account of the false belief transition is incomplete in important ways. After all, our agent had only to choose the best of two known models. This begs an understanding of the dynamics of rational revision near threshold and when the space of possible models is far larger. Further, a single formal model ought ultimately to be applicable to many false belief tasks, and to reasoning about mental states more generally. Several components seem necessary to extend a particular theory of mind into such a framework theory: a richer representation for the propositional content and attitudes in these tasks, extension of the implicit quantifier over trials to one over situations and people, and a broader view of the probability distributions relating mental state variables. Each of these is an important direction for future research.”

“Intuitive Theories of Mind: A Rational Approach to False Belief”
Goodman et al.

“The present account of the false belief transition is incomplete in important ways. After all, our agent had only to choose the best of two known models. This begs an understanding of the dynamics of rational revision near threshold and when the space of possible models is far larger. Further, a single formal model ought ultimately to be applicable to many false belief tasks, and to reasoning about mental states more generally. Several components seem necessary to extend a particular theory of mind into such a framework theory: a richer representation for the propositional content and attitudes in these tasks, extension of the implicit quantifier over trials to one over situations and people, and a broader view of the probability distributions relating mental state variables. Each of these is an important direction for future research.”

“Intuitive Theories of Mind: A Rational Approach to False Belief”
Goodman et al.

Done.

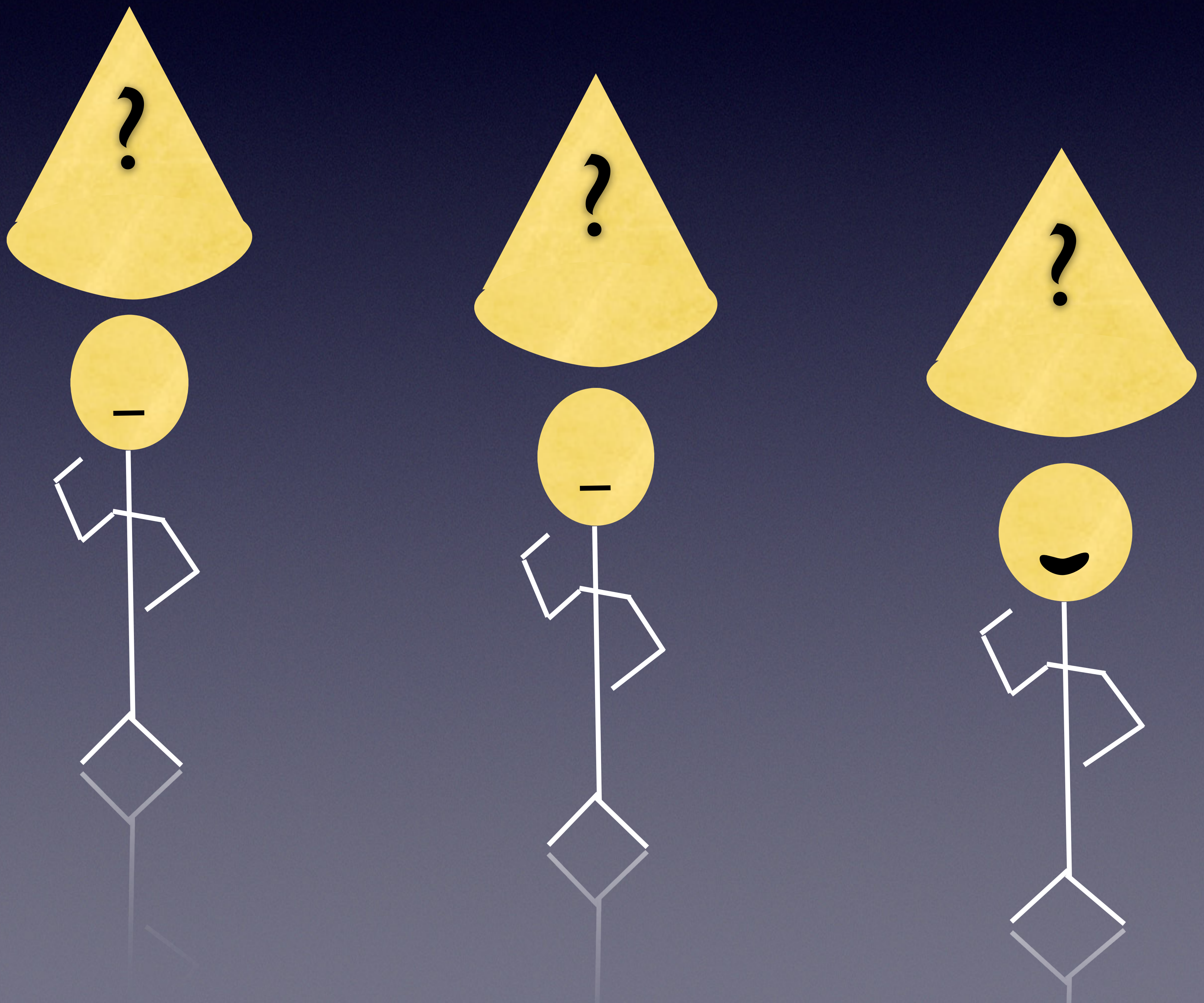
“The present account of the false belief transition is incomplete in important ways. After all, our agent had only to choose the best of two known models. This begs an understanding of the dynamics of rational revision near threshold and when the space of possible models is far larger. Further, a single formal model ought ultimately to be applicable to many false belief tasks, and to reasoning about mental states more generally. Several components seem necessary to extend a particular theory of mind into such a framework theory: a richer representation for the propositional content and attitudes in these tasks, extension of the implicit quantifier over trials to one over situations and people, and a broader view of the probability distributions relating mental state variables. Each of these is an important direction for future research.”

“Intuitive Theories of Mind: A Rational Approach to False Belief”
Goodman et al.

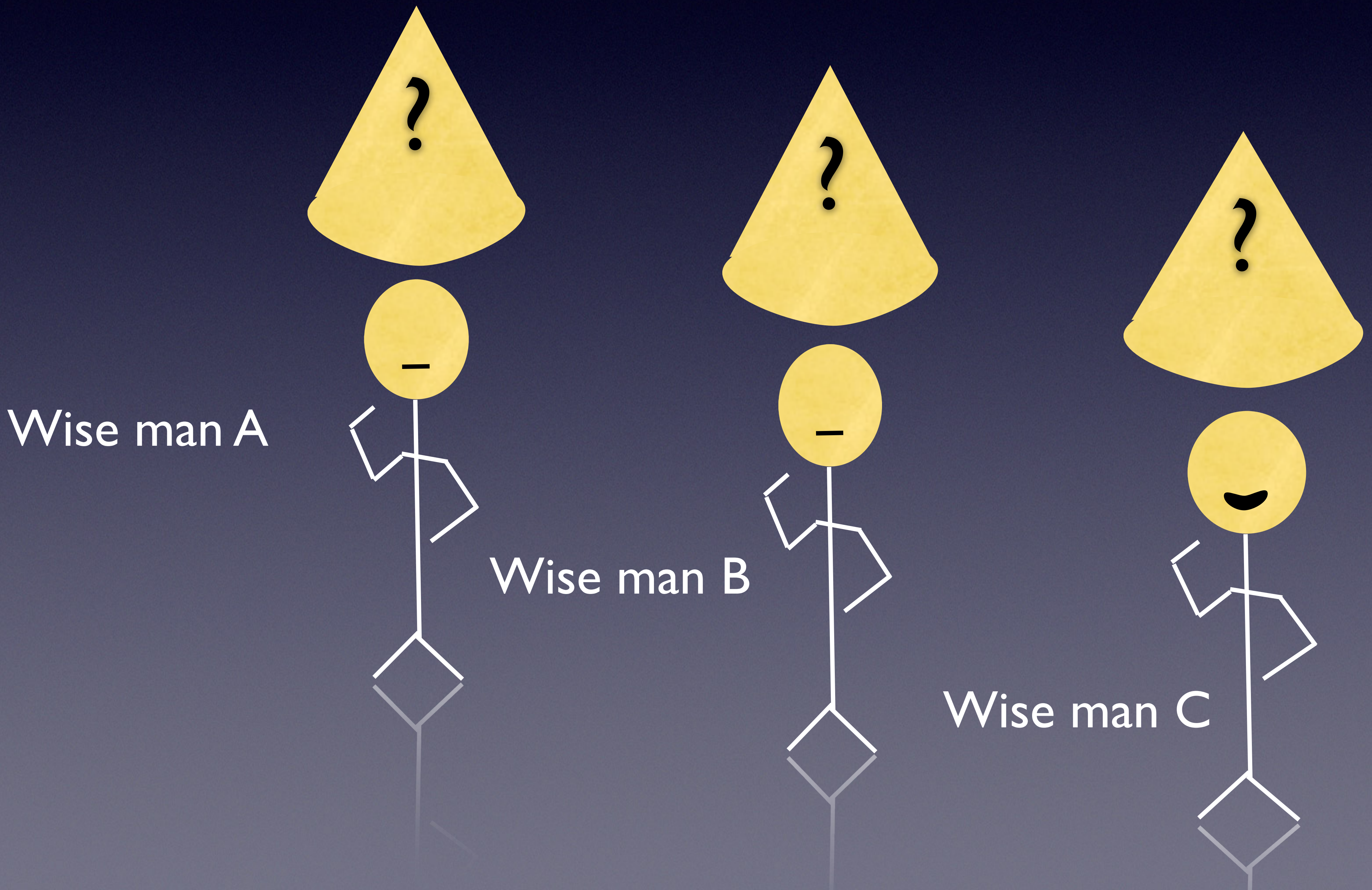
Done.

Cracking Wise Man Tests ...

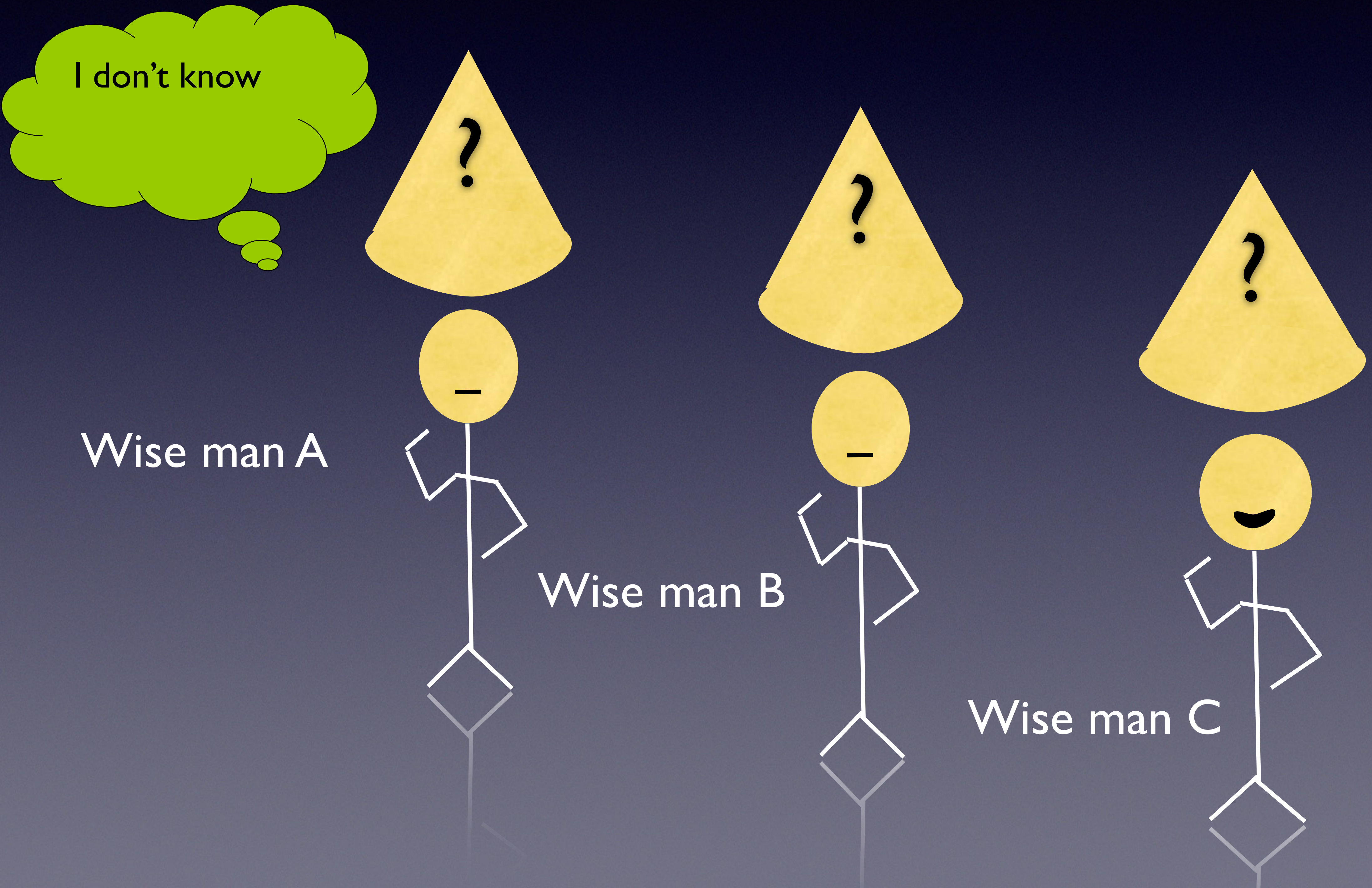
Wise Men Puzzle



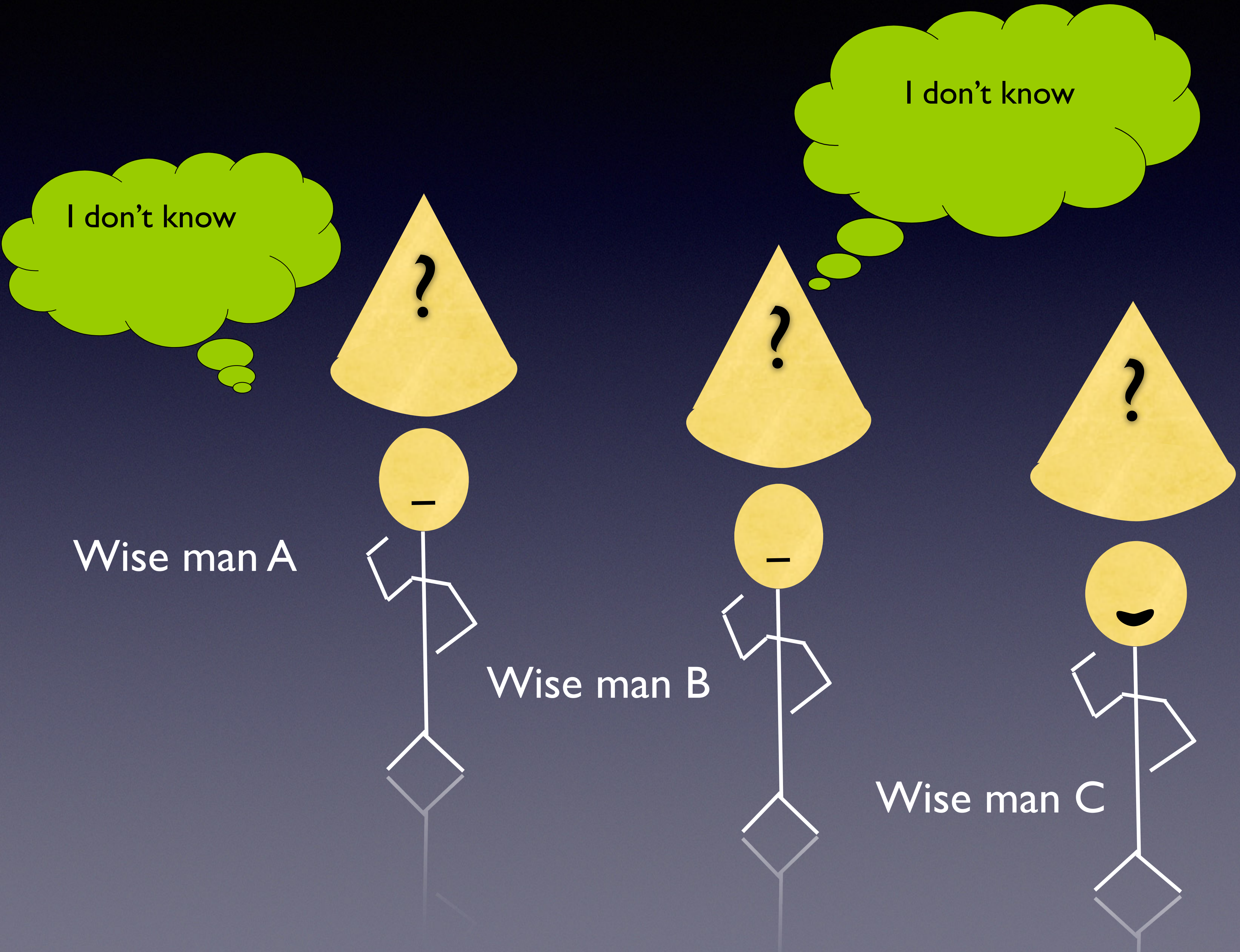
Wise Men Puzzle



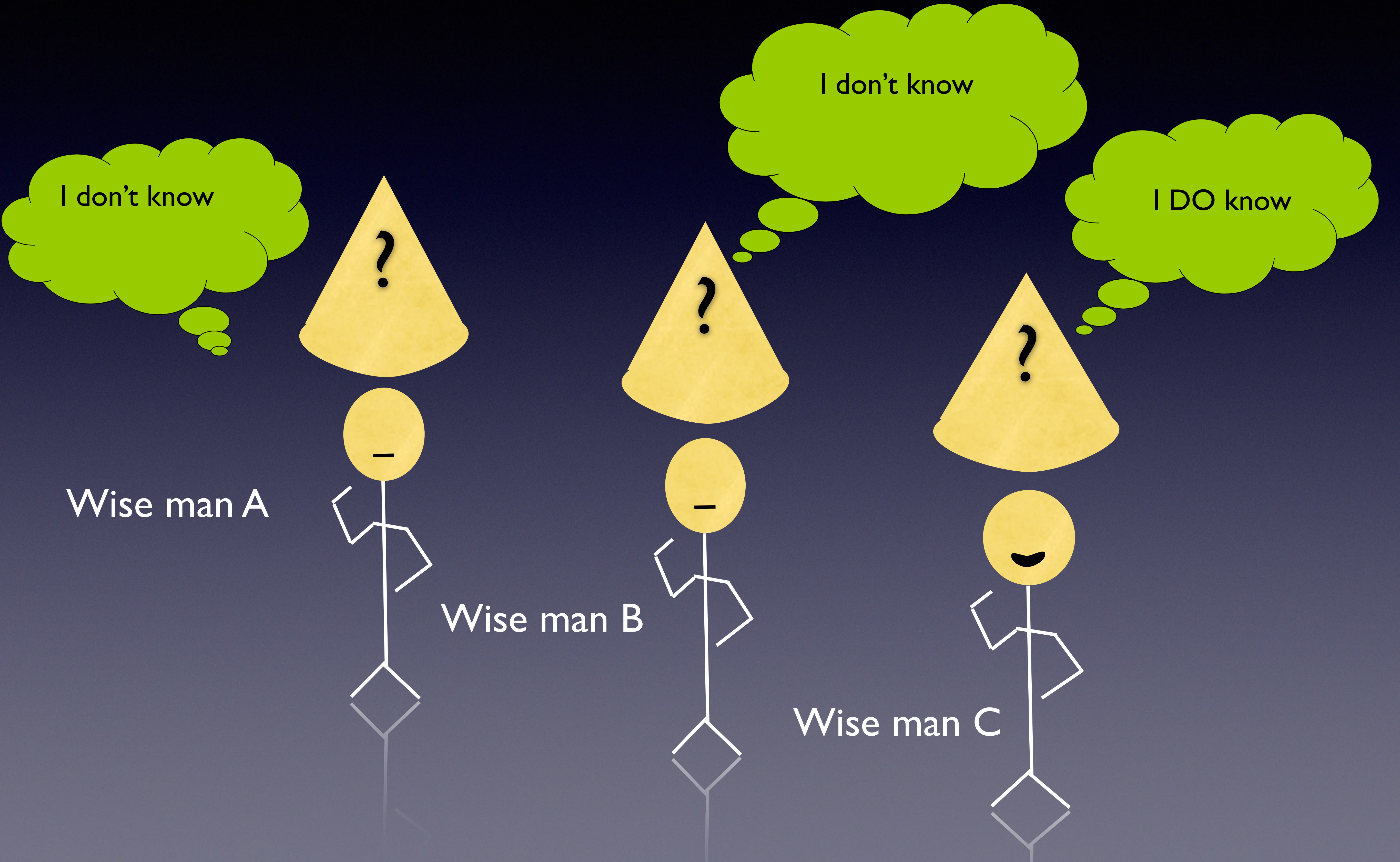
Wise Men Puzzle



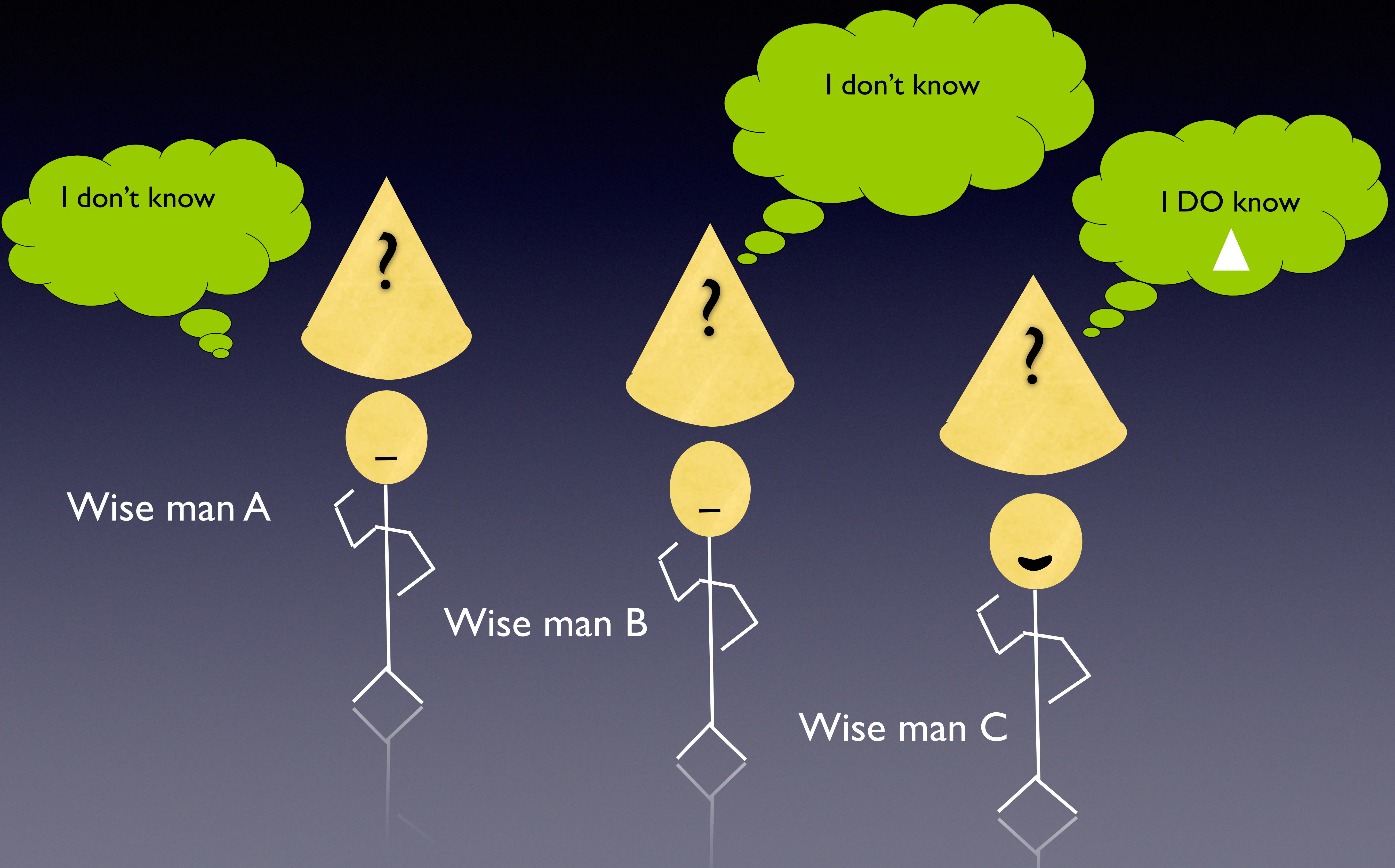
Wise Men Puzzle



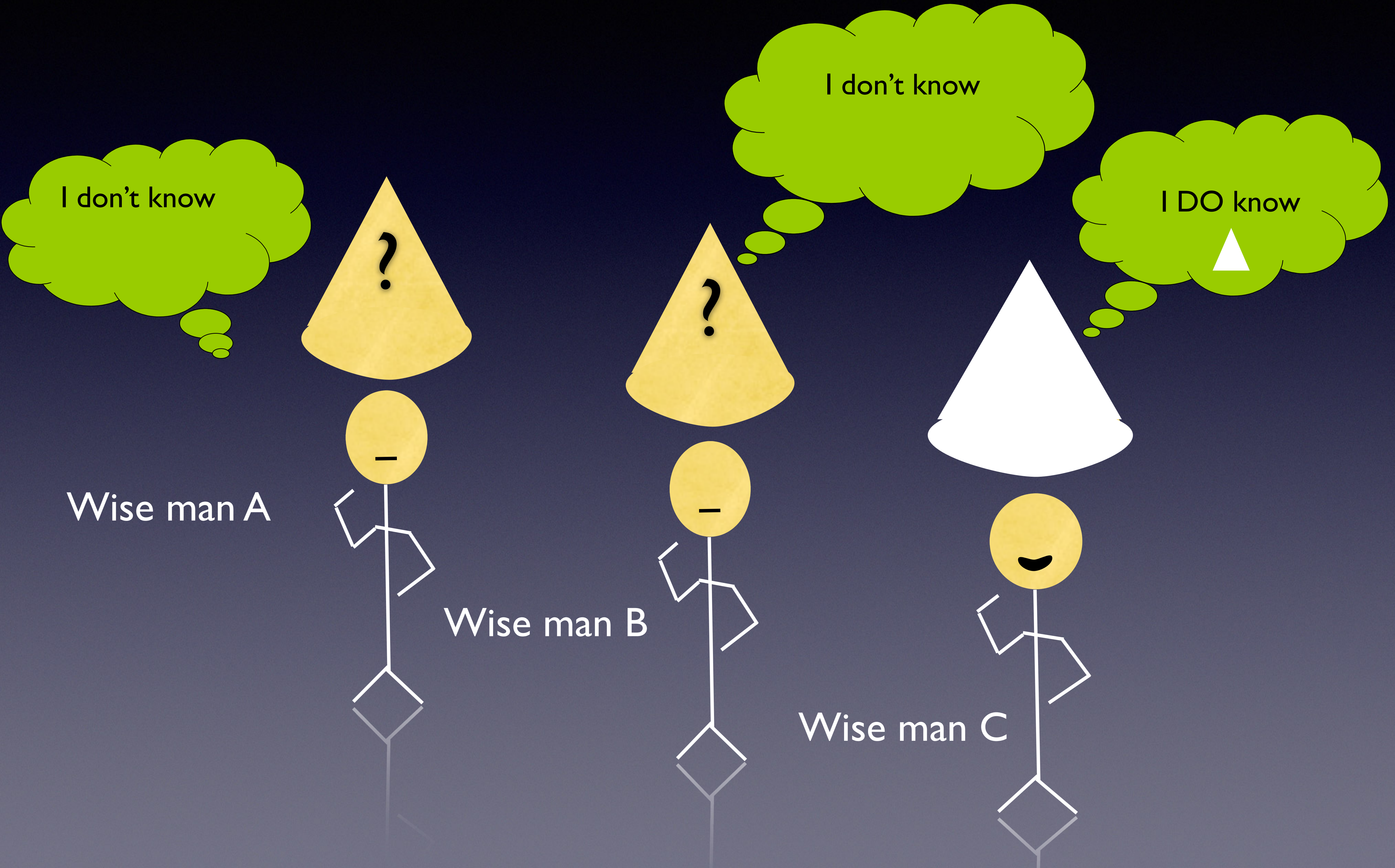
Wise Men Puzzle



Wise Men Puzzle



Wise Men Puzzle



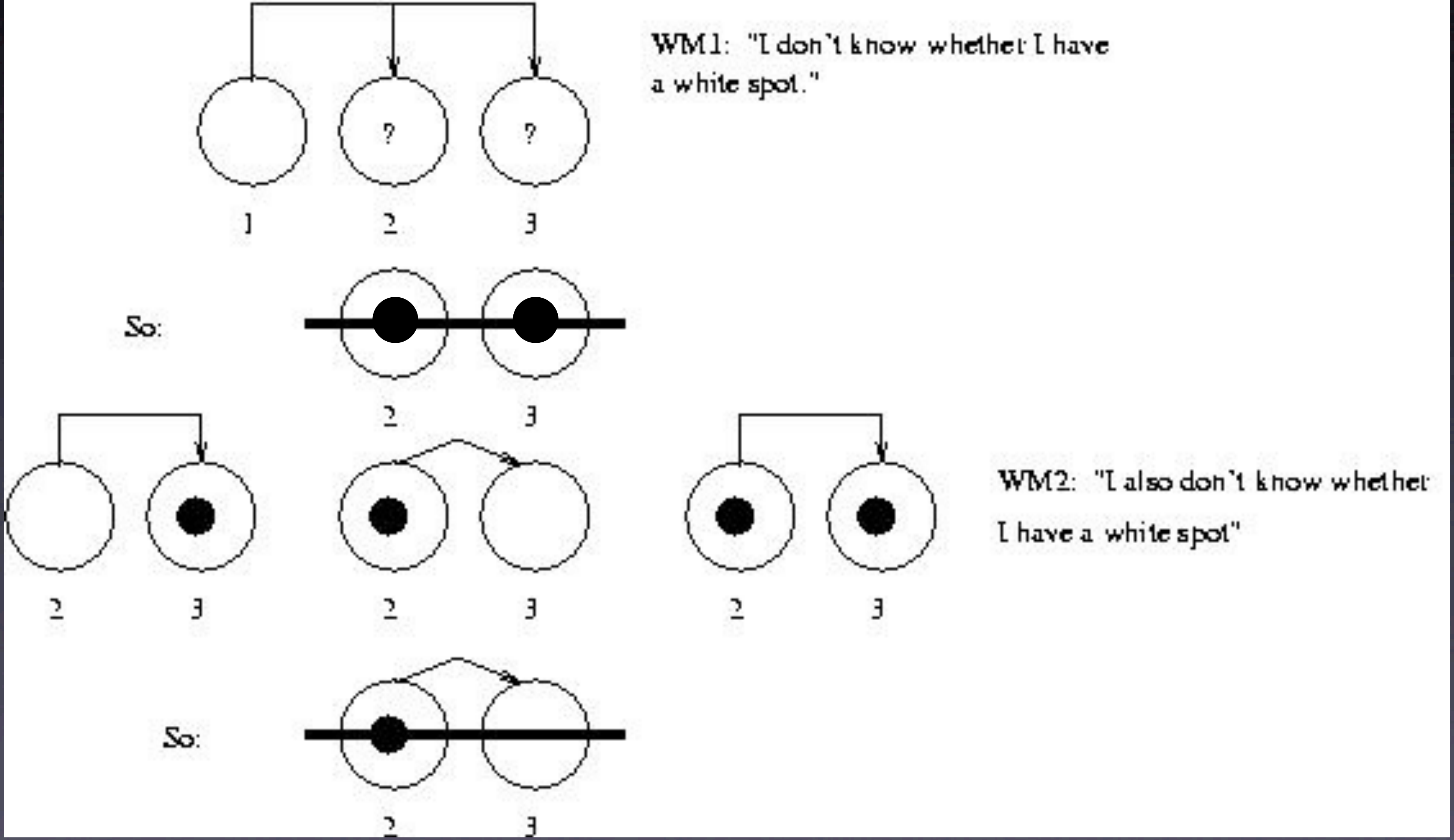
Proof from WM3

Proposition: I have a white fez.

Proof: Remember as a first fact that we all know that at least one of us has a white fez. When the first wise man says that he doesn't know, I immediately know that either WM2 has a white fez, or I do, or both of us does. I know this because if neither of us has a whilte fez, WM1 would have said immediately that in light of our first fact, he has a white fez. My next piece of info comes from what WM2 says; he says that he is *also* ignorant. Now, if he had seen no white fez on my head, he would have immediately said "I have a white fez!" (He would have said this because after WM1 spoke, he carried out the same reasoning I did, and hence ruled out the (WM2-bf & WM3-bf) permutation.) But this *isn't* what he said. Hence, I do have a white fez on my head. QED

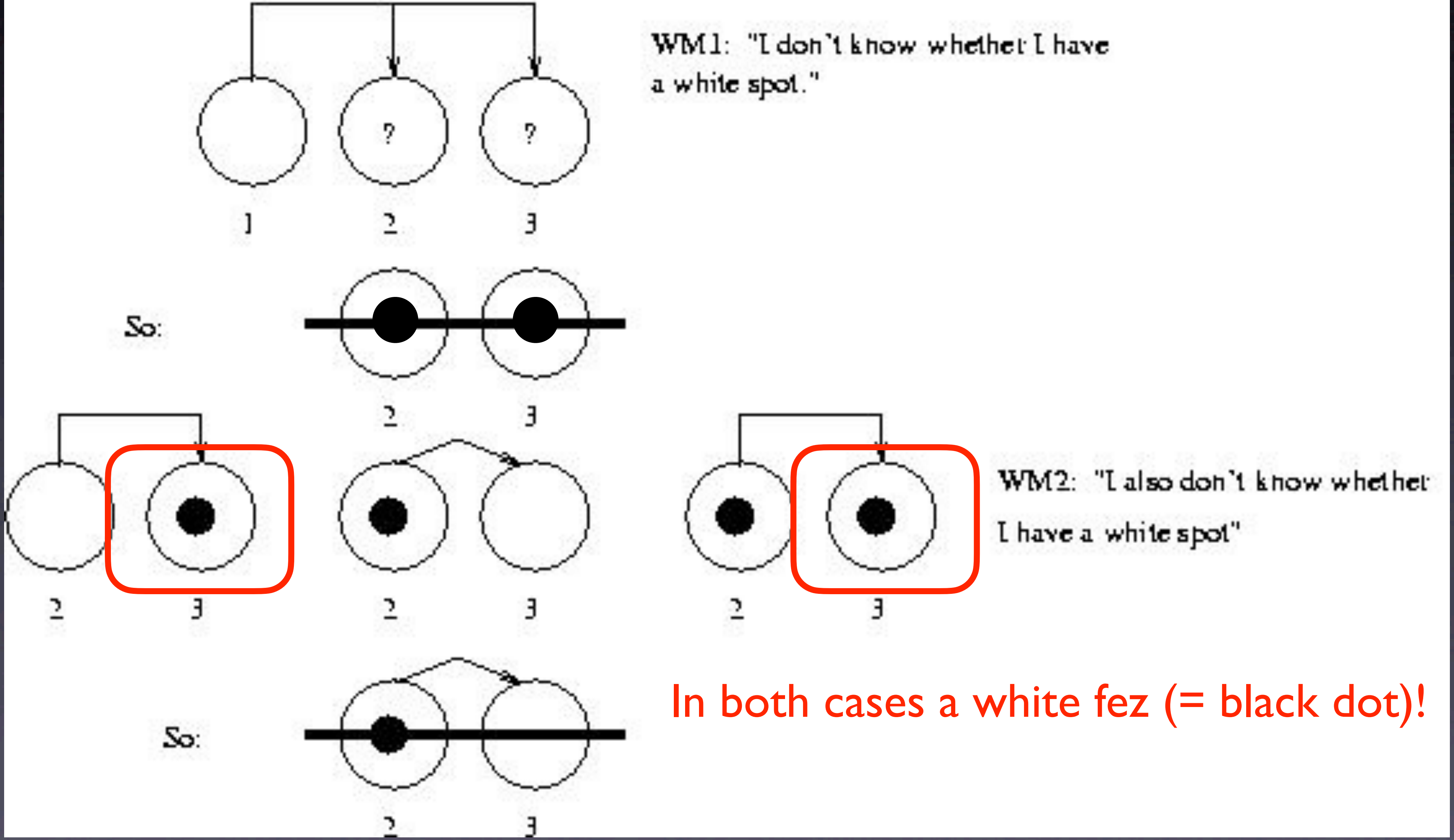
Diagrammatic Version of Reasoning in WMP$_3$

(pov of *truly* wise man; easy for rational humans)


WM1: "I don't know whether I have
a white spot."
?
?
1
2
3
So:
2
3
WM2: "I also don't know whether
I have a white spot"
2
3
2
3
2
3
So:
2
3

Diagrammatic Version of Reasoning in WMP$_3$

(pov of *truly* wise man; easy for rational humans)

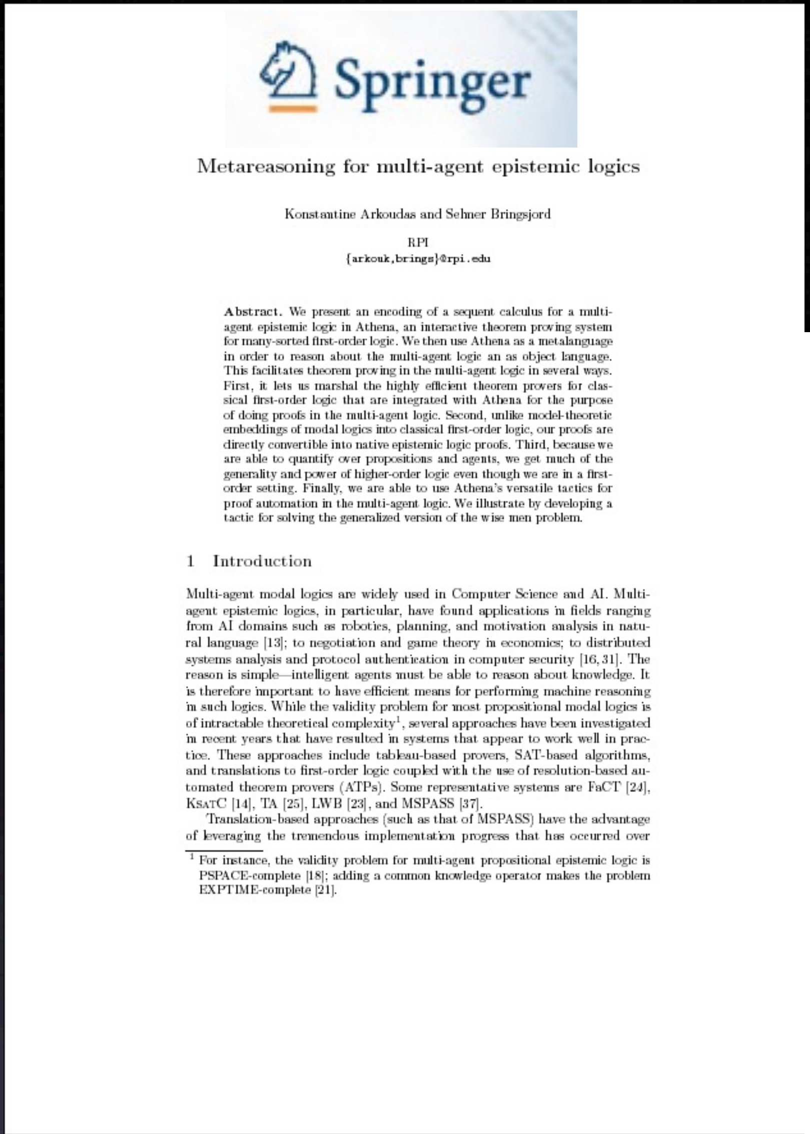


WM1: "I don't know whether I have a white spot."

WM2: "I also don't know whether I have a white spot"

In both cases a white fez (= black dot)!

Arkoudas-Proved-Sound Algorithm for Generating Proof-Theoretic Solution to WMP$_n$



Metareasoning for multi-agent epistemic logics

Konstantine Arkoudas and Selmer Bringsjord

RPI
{arkouk,brings}@rpi.edu

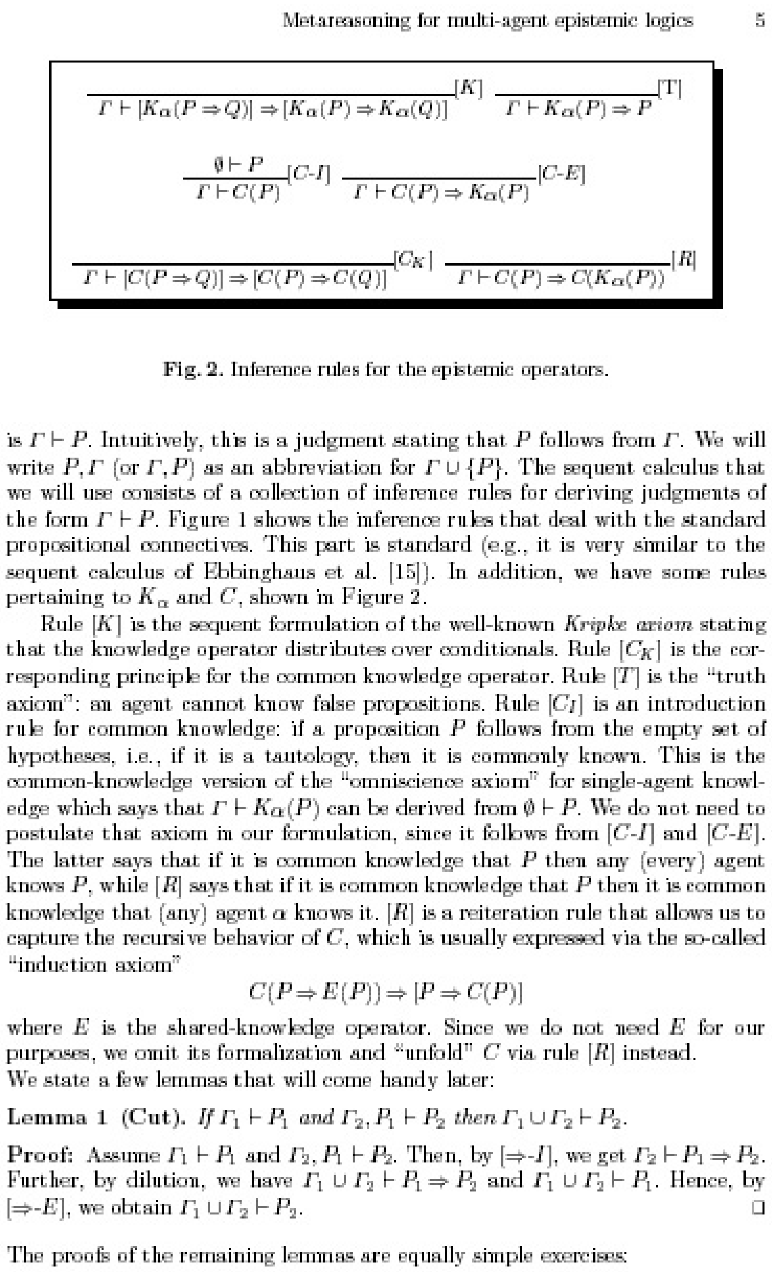
Abstract. We present an encoding of a sequent calculus for a multi-agent epistemic logic in Athena, an interactive theorem proving system for many-sorted first-order logic. We then use Athena as a metalanguage in order to reason about the multi-agent logic an as object language. This facilitates theorem proving in the multi-agent logic in several ways. First, it lets us marshal the highly efficient theorem provers for classical first-order logic that are integrated with Athena for the purpose of doing proofs in the multi-agent logic. Second, unlike model-theoretic embeddings of modal logics into classical first-order logic, our proofs are directly convertible into native epistemic logic proofs. Third, because we are able to quantify over propositions and agents, we get much of the generality and power of higher-order logic even though we are in a first-order setting. Finally, we are able to use Athena's versatile tactics for proof automation in the multi-agent logic. We illustrate by developing a tactic for solving the generalized version of the wise men problem.

1 Introduction

Multi-agent modal logics are widely used in Computer Science and AI. Multi-agent epistemic logics, in particular, have found applications in fields ranging from AI domains such as robotics, planning, and motivation analysis in natural language [13]; to negotiation and game theory in economics; to distributed systems analysis and protocol authentication in computer security [16, 31]. The reason is simple—intelligent agents must be able to reason about knowledge. It is therefore important to have efficient means for performing machine reasoning in such logics. While the validity problem for most propositional modal logics is of intractable theoretical complexity[1], several approaches have been investigated in recent years that have resulted in systems that appear to work well in practice. These approaches include tableau-based provers, SAT-based algorithms, and translations to first-order logic coupled with the use of resolution-based automated theorem provers (ATPs). Some representative systems are FaCT [24], KsatC [14], TA [25], LWB [23], and MSPASS [37].

Translation-based approaches (such as that of MSPASS) have the advantage of leveraging the tremendous implementation progress that has occurred over

[1] For instance, the validity problem for multi-agent propositional epistemic logic is PSPACE-complete [18]; adding a common knowledge operator makes the problem EXPTIME-complete [21].



$$\frac{}{\Gamma \vdash [K_\alpha(P \Rightarrow Q)] \Rightarrow [K_\alpha(P) \Rightarrow K_\alpha(Q)]}[K] \qquad \frac{}{\Gamma \vdash K_\alpha(P) \Rightarrow P}[\mathrm{T}]$$

$$\frac{\emptyset \vdash P}{\Gamma \vdash C(P)}[C\text{-}I] \qquad \frac{}{\Gamma \vdash C(P) \Rightarrow K_\alpha(P)}[C\text{-}E]$$

$$\frac{}{\Gamma \vdash [C(P \Rightarrow Q)] \Rightarrow [C(P) \Rightarrow C(Q)]}[C_K] \qquad \frac{}{\Gamma \vdash C(P) \Rightarrow C(K_\alpha(P))}[R]$$

Fig. 2. Inference rules for the epistemic operators.

is $\Gamma \vdash P$. Intuitively, this is a judgment stating that P follows from Γ. We will write P, Γ (or Γ, P) as an abbreviation for $\Gamma \cup \{P\}$. The sequent calculus that we will use consists of a collection of inference rules for deriving judgments of the form $\Gamma \vdash P$. Figure 1 shows the inference rules that deal with the standard propositional connectives. This part is standard (e.g., it is very similar to the sequent calculus of Ebbinghaus et al. [15]). In addition, we have some rules pertaining to K_α and C, shown in Figure 2.

Rule $[K]$ is the sequent formulation of the well-known *Kripke axiom* stating that the knowledge operator distributes over conditionals. Rule $[C_K]$ is the corresponding principle for the common knowledge operator. Rule $[T]$ is the "truth axiom": an agent cannot know false propositions. Rule $[C_I]$ is an introduction rule for common knowledge: if a proposition P follows from the empty set of hypotheses, i.e., if it is a tautology, then it is commonly known. This is the common-knowledge version of the "omniscience axiom" for single-agent knowledge which says that $\Gamma \vdash K_\alpha(P)$ can be derived from $\emptyset \vdash P$. We do not need to postulate that axiom in our formulation, since it follows from $[C\text{-}I]$ and $[C\text{-}E]$. The latter says that if it is common knowledge that P then any (every) agent knows P, while $[R]$ says that if it is common knowledge that P then it is common knowledge that (any) agent α knows it. $[R]$ is a reiteration rule that allows us to capture the recursive behavior of C, which is usually expressed via the so-called "induction axiom"

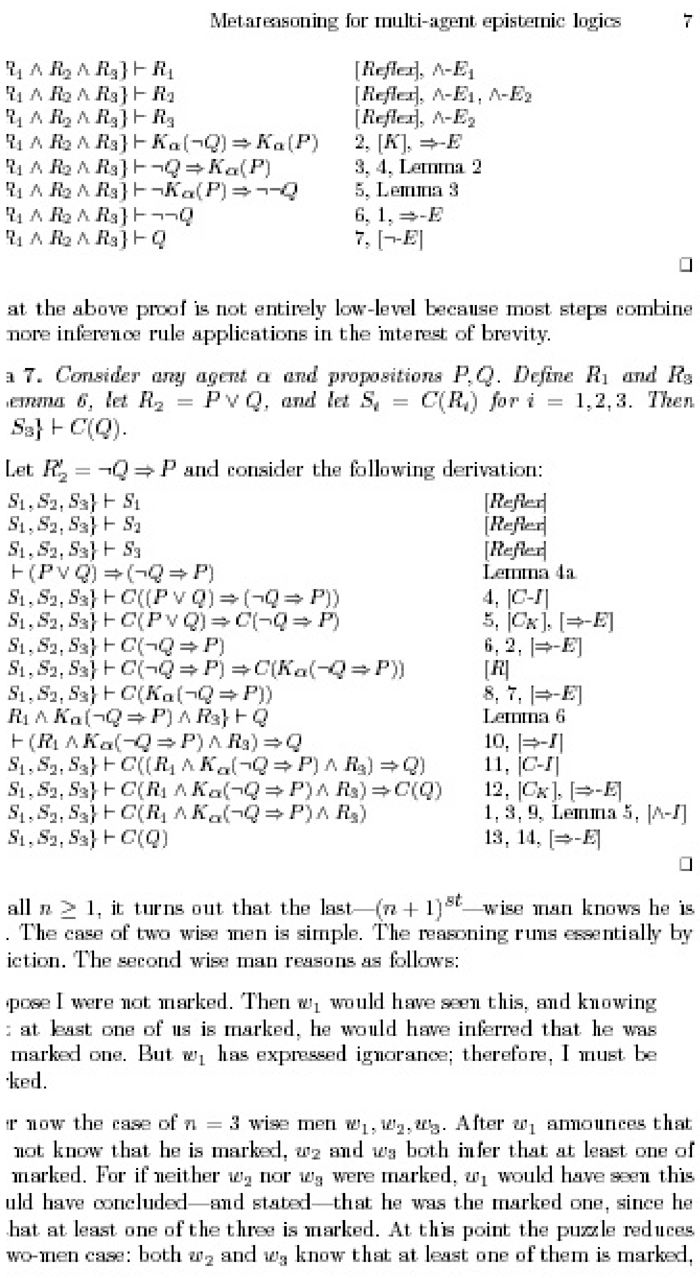
$$C(P \Rightarrow E(P)) \Rightarrow [P \Rightarrow C(P)]$$

where E is the shared-knowledge operator. Since we do not need E for our purposes, we omit its formalization and "unfold" C via rule $[R]$ instead.
We state a few lemmas that will come handy later:

Lemma 1 (Cut). *If $\Gamma_1 \vdash P_1$ and $\Gamma_2, P_1 \vdash P_2$ then $\Gamma_1 \cup \Gamma_2 \vdash P_2$.*

Proof: Assume $\Gamma_1 \vdash P_1$ and $\Gamma_2, P_1 \vdash P_2$. Then, by $[\Rightarrow\text{-}I]$, we get $\Gamma_2 \vdash P_1 \Rightarrow P_2$. Further, by dilution, we have $\Gamma_1 \cup \Gamma_2 \vdash P_1 \Rightarrow P_2$ and $\Gamma_1 \cup \Gamma_2 \vdash P_1$. Hence, by $[\Rightarrow\text{-}E]$, we obtain $\Gamma_1 \cup \Gamma_2 \vdash P_2$. □

The proofs of the remaining lemmas are equally simple exercises:



$R_1 \wedge R_2 \wedge R_3\} \vdash R_1$	$[Reflex]$, $\wedge\text{-}E_1$
$R_1 \wedge R_2 \wedge R_3\} \vdash R_2$	$[Reflex]$, $\wedge\text{-}E_1$, $\wedge\text{-}E_2$
$R_1 \wedge R_2 \wedge R_3\} \vdash R_3$	$[Reflex]$, $\wedge\text{-}E_2$
$R_1 \wedge R_2 \wedge R_3\} \vdash K_\alpha(\neg Q) \Rightarrow K_\alpha(P)$	2, $[K]$, $\Rightarrow\text{-}E$
$R_1 \wedge R_2 \wedge R_3\} \vdash \neg Q \Rightarrow K_\alpha(P)$	3, 4, Lemma 2
$R_1 \wedge R_2 \wedge R_3\} \vdash \neg K_\alpha(P) \Rightarrow \neg\neg Q$	5, Lemma 3
$R_1 \wedge R_2 \wedge R_3\} \vdash \neg\neg Q$	6, 1, $\Rightarrow\text{-}E$
$R_1 \wedge R_2 \wedge R_3\} \vdash Q$	7, $[\neg\text{-}E]$

□

at the above proof is not entirely low-level because most steps combine nore inference rule applications in the interest of brevity.

a 7. *Consider any agent α and propositions P, Q. Define R_1 and R_3 emma 6, let $R_2 = P \vee Q$, and let $S_i = C(R_i)$ for $i = 1, 2, 3$. Then $S_3\} \vdash C(Q)$.*

Let $R_2' = \neg Q \Rightarrow P$ and consider the following derivation:

$S_1, S_2, S_3\} \vdash S_1$	$[Reflex]$
$S_1, S_2, S_3\} \vdash S_2$	$[Reflex]$
$S_1, S_2, S_3\} \vdash S_3$	$[Reflex]$
$\vdash (P \vee Q) \Rightarrow (\neg Q \Rightarrow P)$	Lemma 4a
$S_1, S_2, S_3\} \vdash C((P \vee Q) \Rightarrow (\neg Q \Rightarrow P))$	4, $[C\text{-}I]$
$S_1, S_2, S_3\} \vdash C(P \vee Q) \Rightarrow C(\neg Q \Rightarrow P)$	5, $[C_K]$, $[\Rightarrow\text{-}E]$
$S_1, S_2, S_3\} \vdash C(\neg Q \Rightarrow P)$	6, 2, $[\Rightarrow\text{-}E]$
$S_1, S_2, S_3\} \vdash C(\neg Q \Rightarrow P) \Rightarrow C(K_\alpha(\neg Q \Rightarrow P))$	$[R]$
$S_1, S_2, S_3\} \vdash C(K_\alpha(\neg Q \Rightarrow P))$	8, 7, $[\Rightarrow\text{-}E]$
$R_1 \wedge K_\alpha(\neg Q \Rightarrow P) \wedge R_3\} \vdash Q$	Lemma 6
$\vdash (R_1 \wedge K_\alpha(\neg Q \Rightarrow P) \wedge R_3) \Rightarrow Q$	10, $[\Rightarrow\text{-}I]$
$S_1, S_2, S_3\} \vdash C((R_1 \wedge K_\alpha(\neg Q \Rightarrow P) \wedge R_3) \Rightarrow Q)$	11, $[C\text{-}I]$
$S_1, S_2, S_3\} \vdash C(R_1 \wedge K_\alpha(\neg Q \Rightarrow P) \wedge R_3) \Rightarrow C(Q)$	12, $[C_K]$, $[\Rightarrow\text{-}E]$
$S_1, S_2, S_3\} \vdash C(R_1 \wedge K_\alpha(\neg Q \Rightarrow P) \wedge R_3)$	1, 3, 9, Lemma 5, $[\wedge\text{-}I]$
$S_1, S_2, S_3\} \vdash C(Q)$	13, 14, $[\Rightarrow\text{-}E]$

□

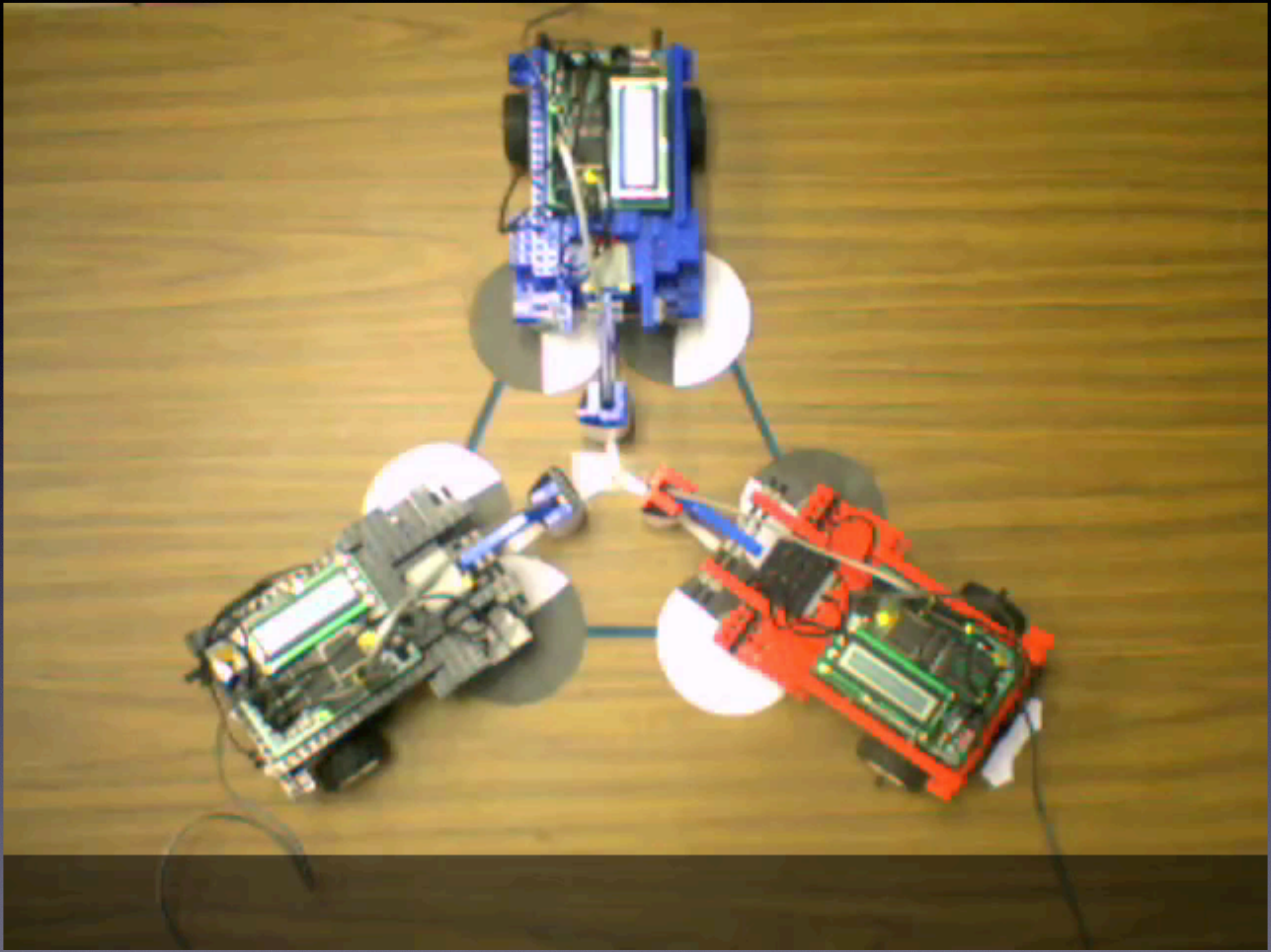
all $n \geq 1$, it turns out that the last—$(n+1)^{st}$—wise man knows he is . The case of two wise men is simple. The reasoning runs essentially by iction. The second wise man reasons as follows:

pose I were not marked. Then w_1 would have seen this, and knowing : at least one of us is marked, he would have inferred that he was marked one. But w_1 has expressed ignorance; therefore, I must be ·ked.

r now the case of $n = 3$ wise men w_1, w_2, w_3. After w_1 announces that not know that he is marked, w_2 and w_3 both infer that at least one of marked. For if neither w_2 nor w_3 were marked, w_1 would have seen this uld have concluded—and stated—that he was the marked one, since he hat at least one of the three is marked. At this point the puzzle reduces wo-men case: both w_2 and w_3 know that at least one of them is marked,

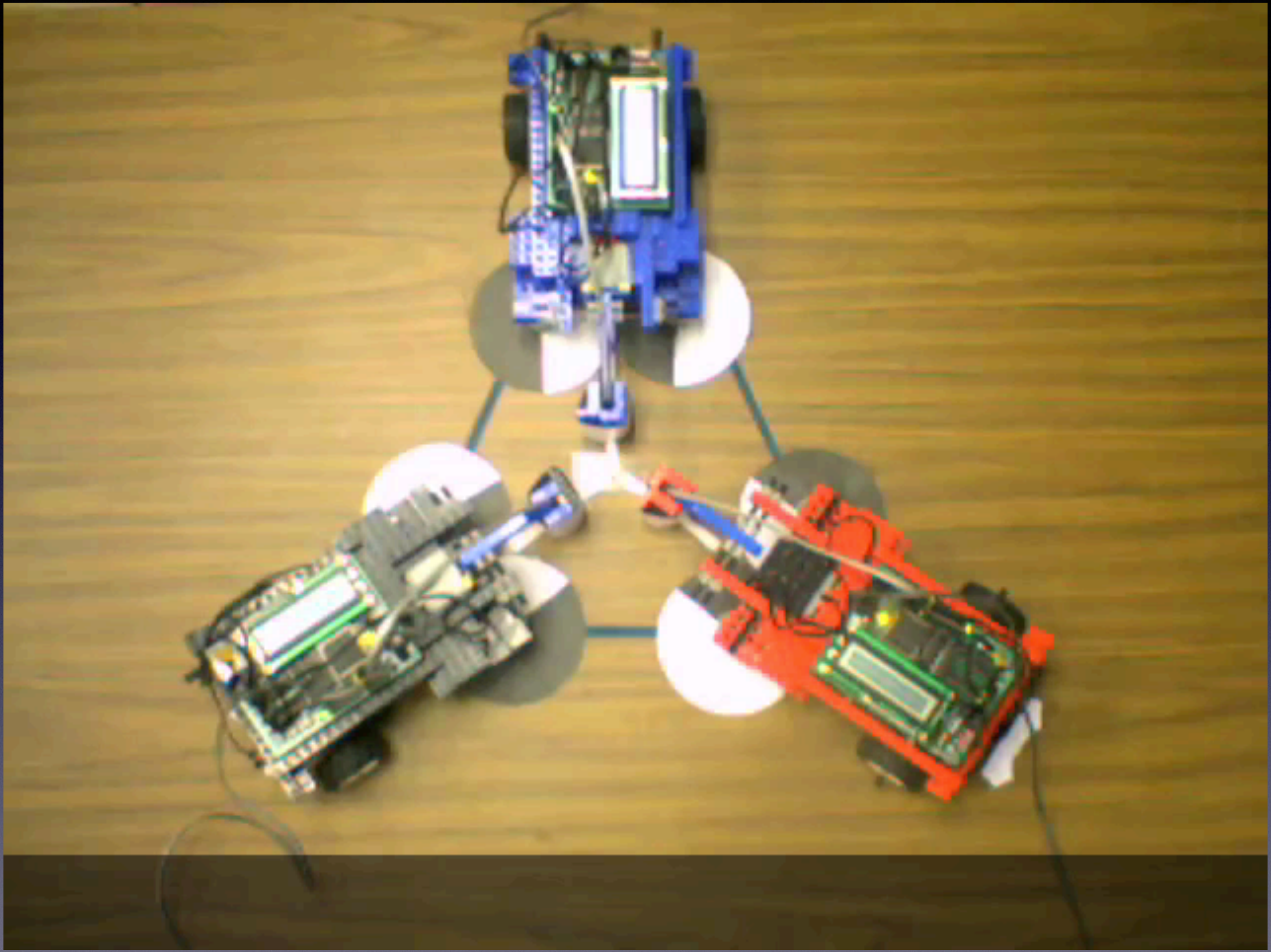
All our human-authored proofs machine-checked.

“Life and Death” Wise Man Test (3)



* Again: Object-level reasoning, reasoning that *produces* object-level reasoning (e.g., methods), and direct, “dirty,” purely computational procedures.

“Life and Death” Wise Man Test (3)



* Again: Object-level reasoning, reasoning that *produces* object-level reasoning (e.g., methods), and direct, “dirty,” purely computational procedures.

Now, harder ...

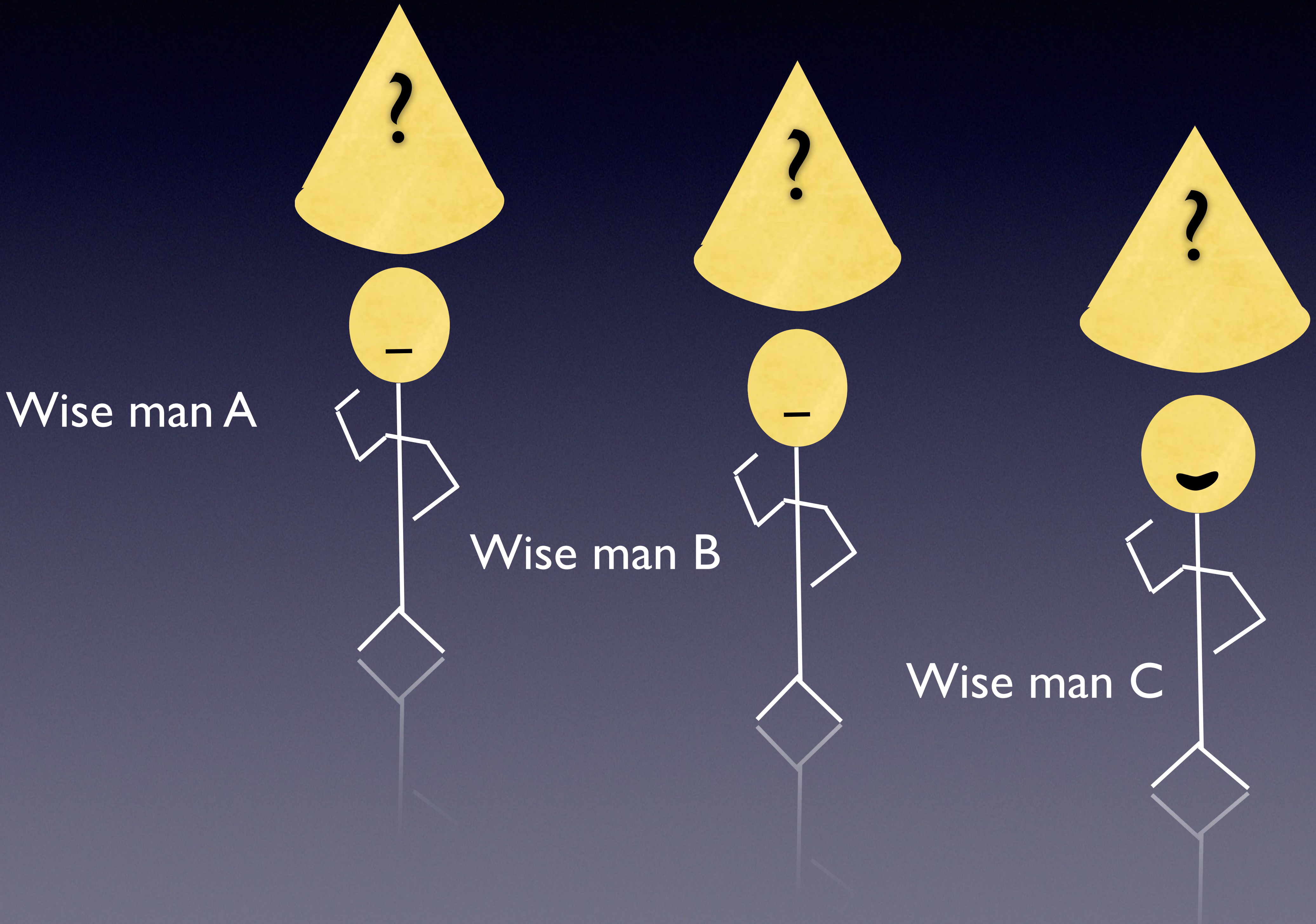
Floridi’s Continuum (augmented), and Claims

	False Belief Task	Wise Man Test (*n*)	Deafening Test	Torture Boots Test	Ultimate Sifter	*Infinitary* False Belief Task
Cutting-Edge AI	Yes	Yes	No	No	No	?
Zombies	Yes	Yes	Yes	Yes	No	?
Human Persons (s-conscious! p-conscious!)	Yes	Yes	Yes	Yes	Yes	Yes

Floridi’s Continuum (augmented), and Claims

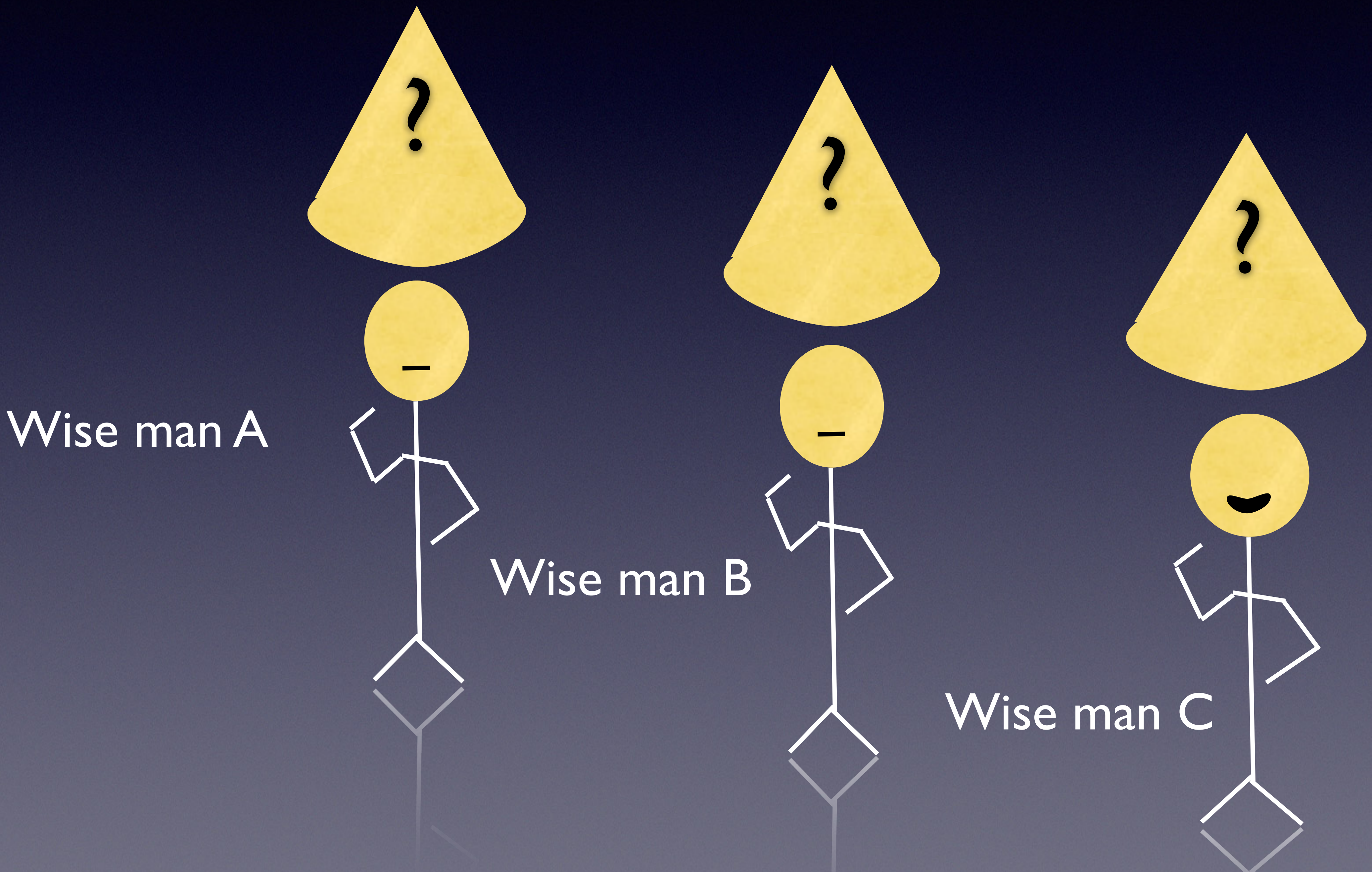
	False Belief Task	Wise Man Test (*n*)	Deafening Test	Torture Boots Test	Ultimate Sifter	*Infinitary* False Belief Task
Cutting-Edge AI	Yes	Yes	No	No	No	?
Zombies	Yes	Yes	Yes	Yes	No	?
Human Persons (s-conscious! p-conscious!)	Yes	Yes	Yes	Yes	Yes	Yes

Floridi's "Ultimate (s- and p-consciousness) Sifter"



poison innocuous

● ● ○ ○ ○

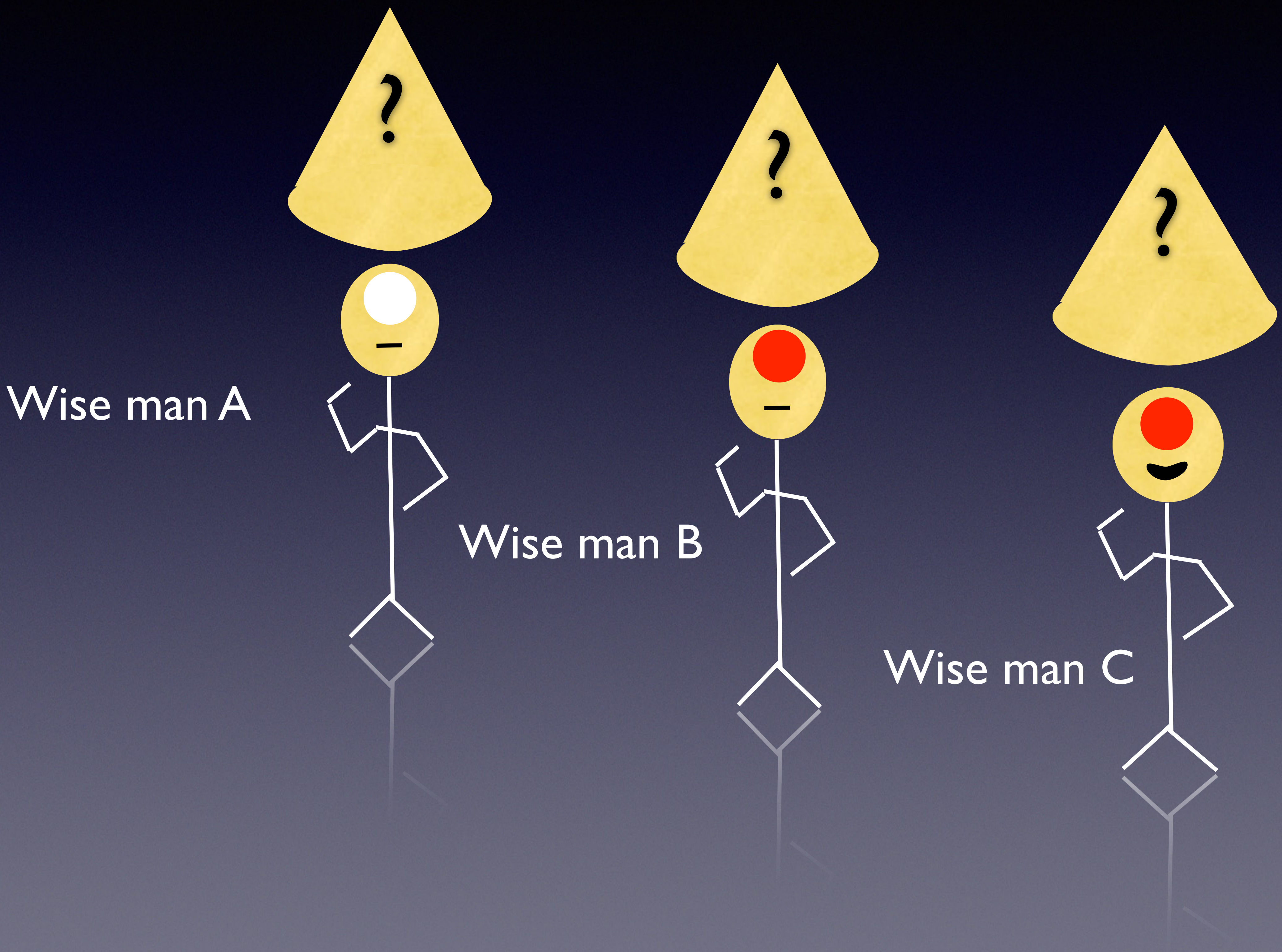


Wise man A

Wise man B

Wise man C

Poison pill strikes the taker dumb.

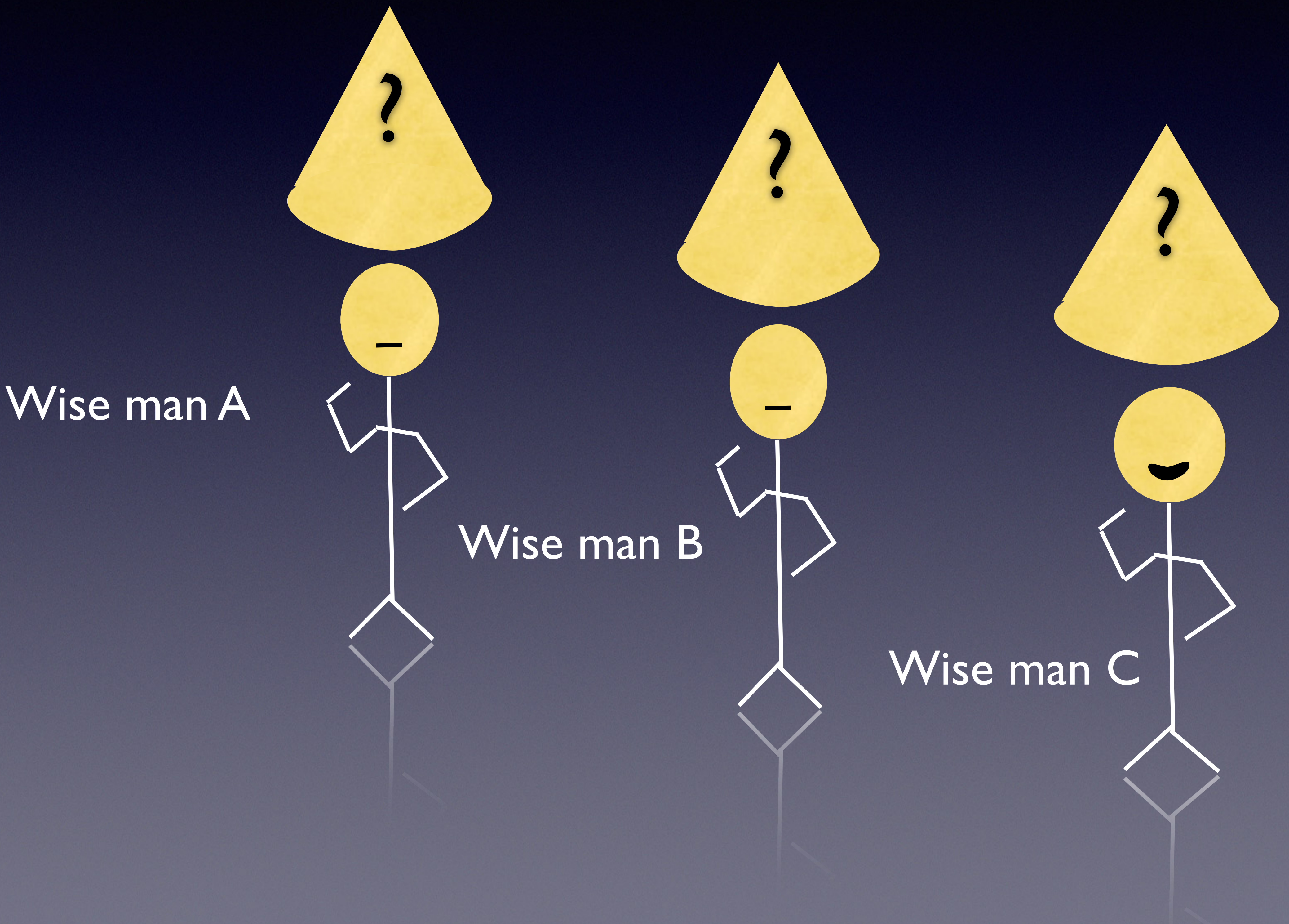


Wise man A

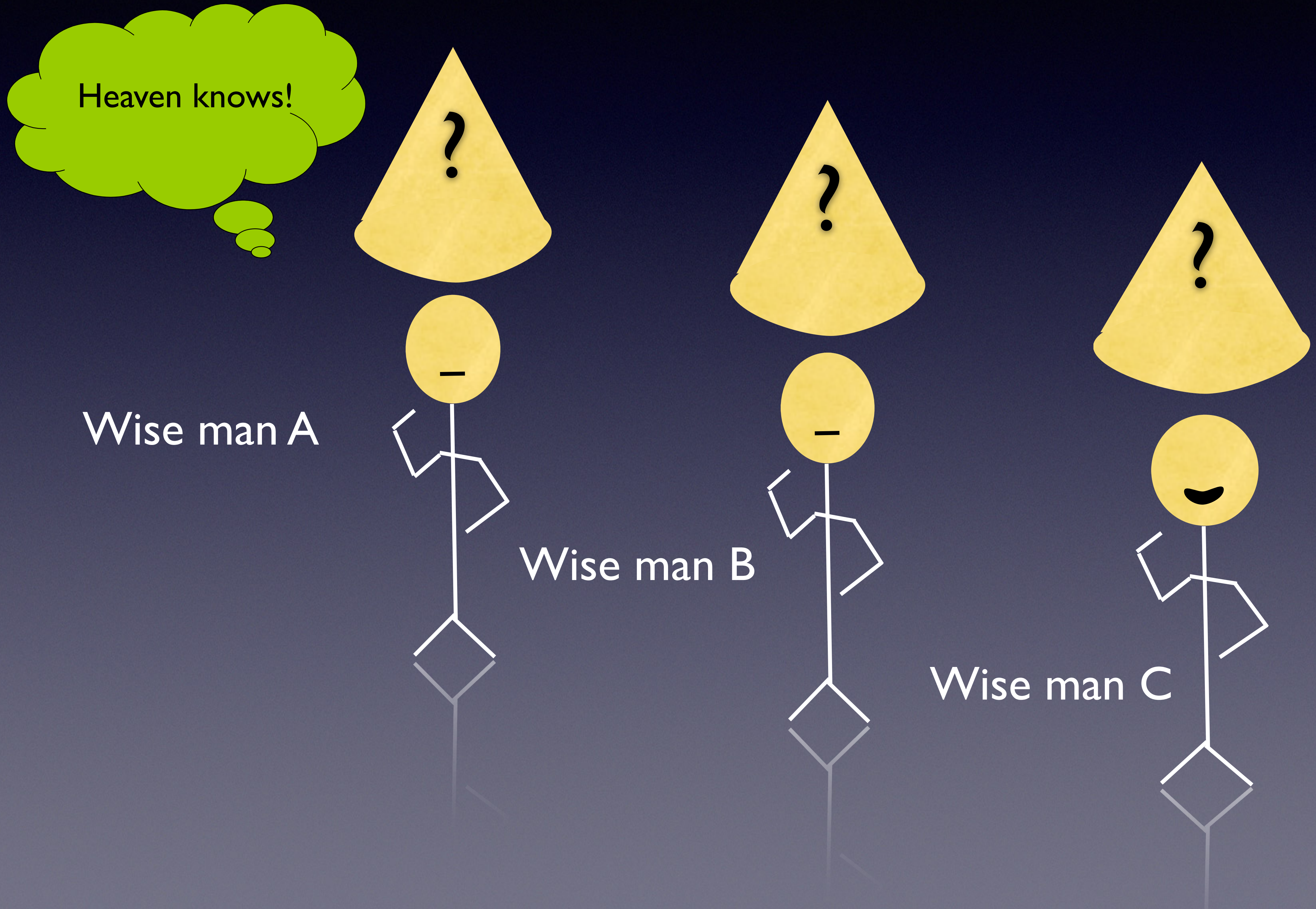
Wise man B

Wise man C

“Have you been struck dumb? As always: Prove it!”



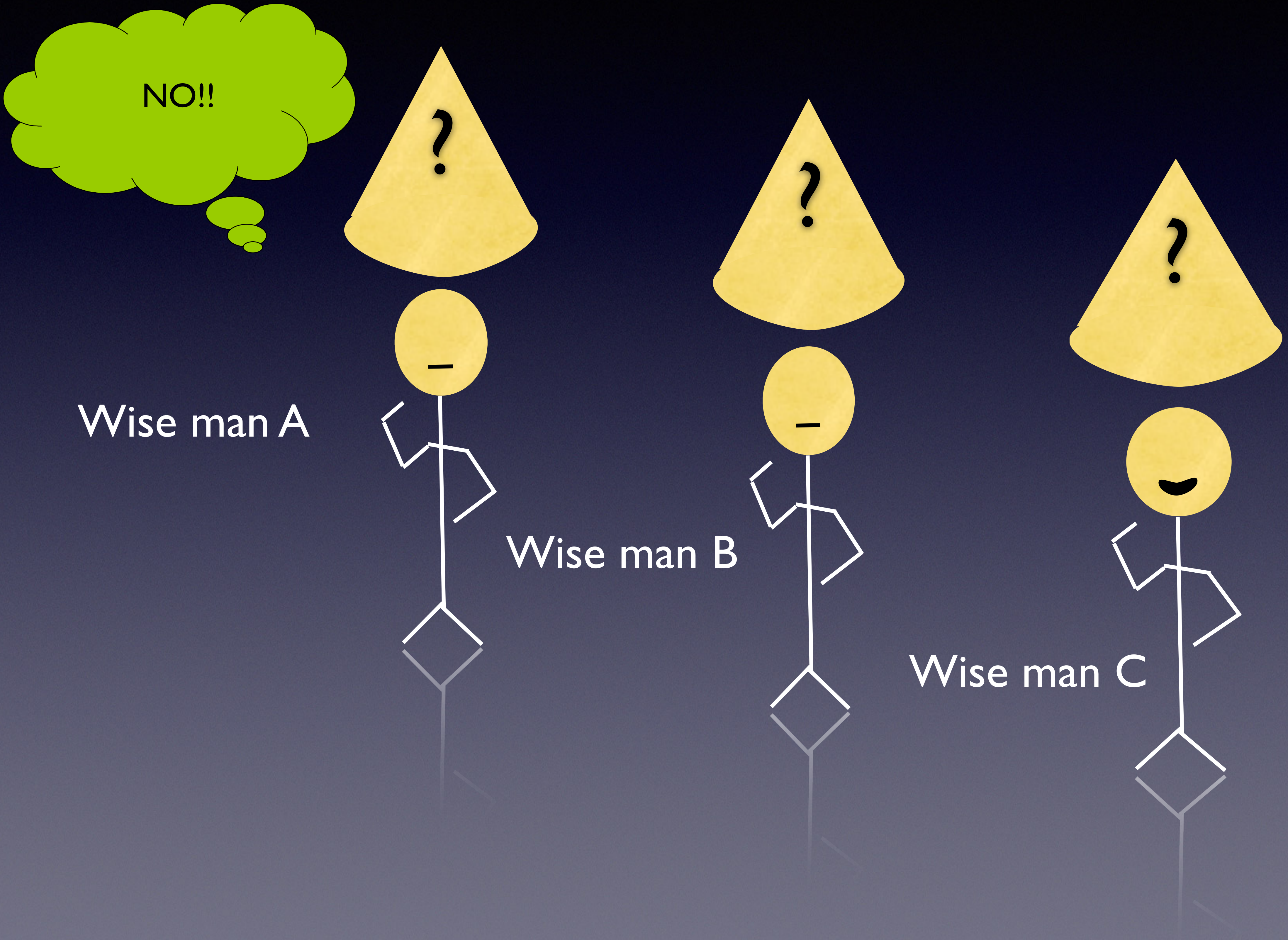
“Have you been struck dumb?”



Two possibilities:

Subsequent silence: failure/death.

Or ...



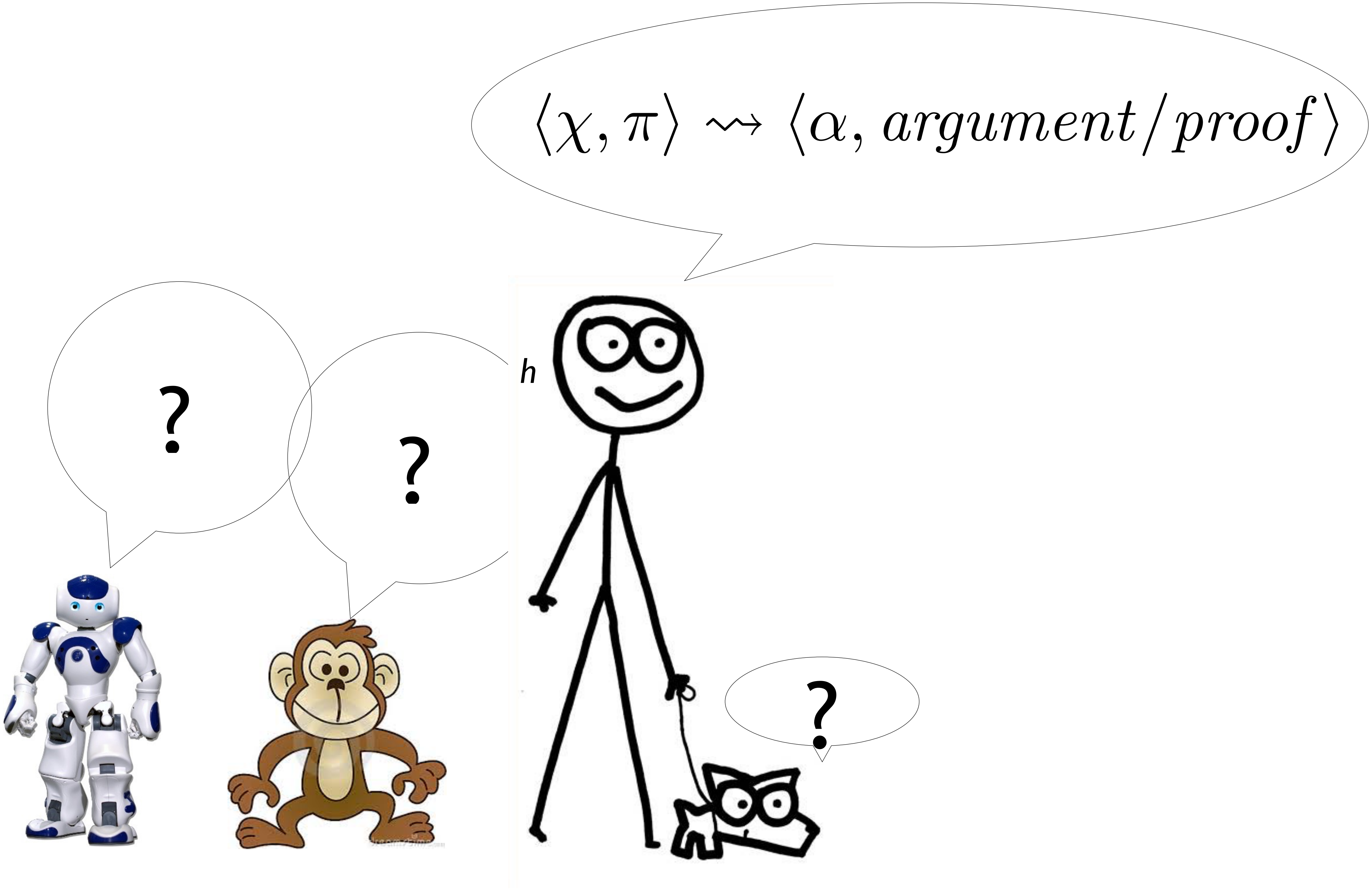
Wise man A

Wise man B

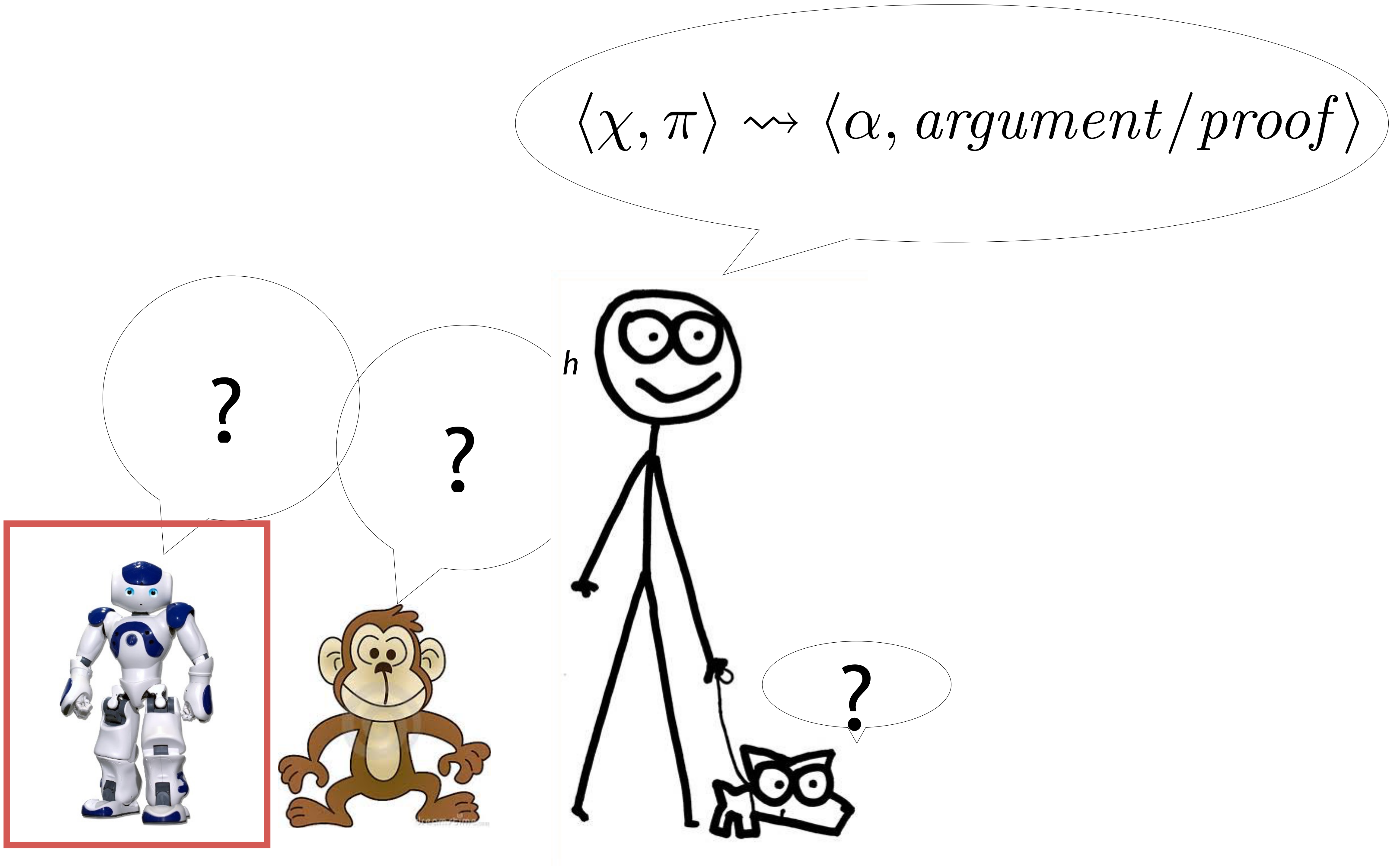
Wise man C

"Had I taken the dumbing tablet I would not have been able to report orally my state of ignorance about my dumb/non-dumb state, but I have been, and I know that I have been, as I have heard myself speaking and saw the guard reacting to my speaking, but this (my oral report) is possible only if I did not take the dumbing tablet, so I know I know I am in the non-dumb state, hence I know that ..."

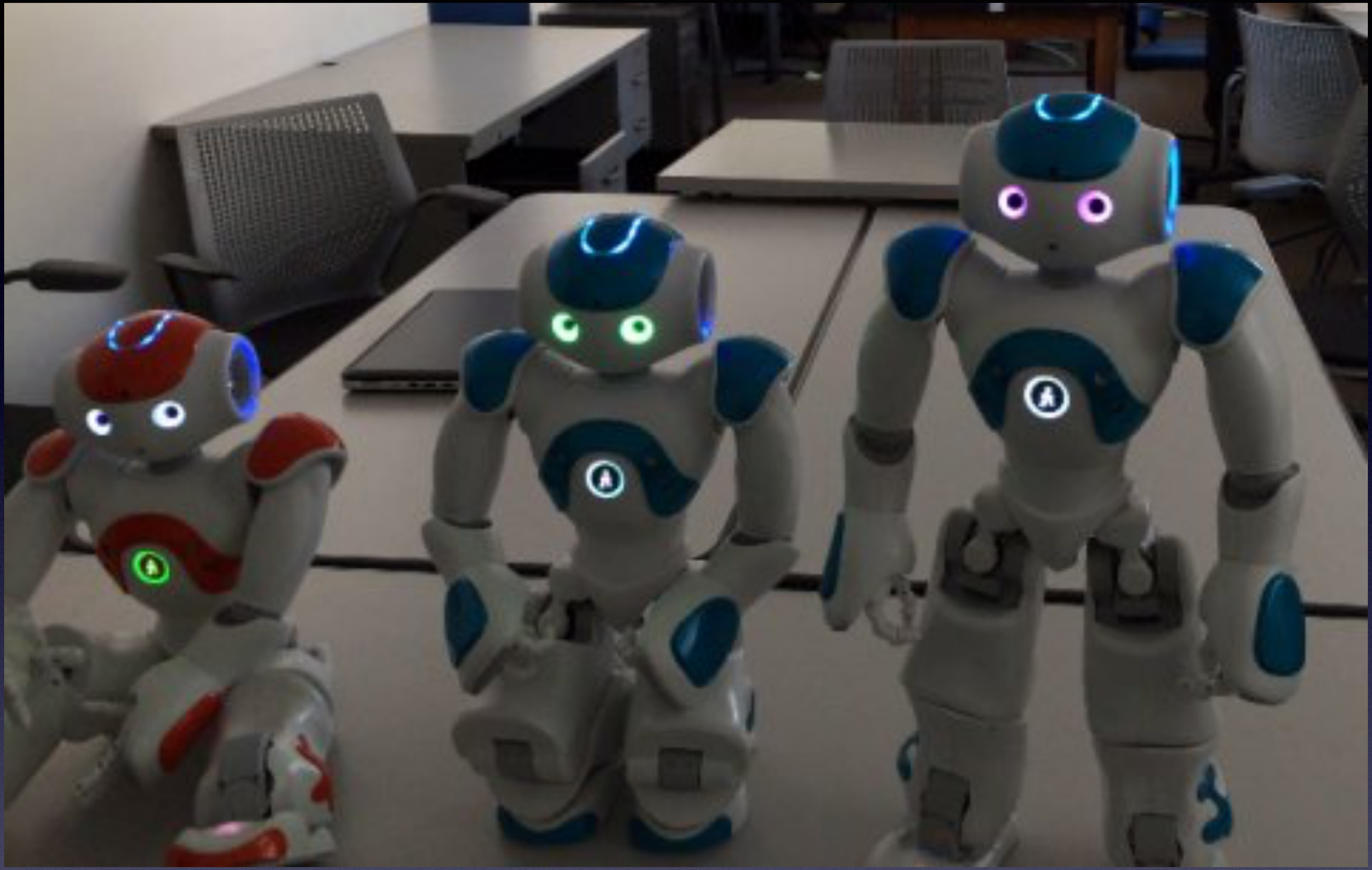
—Luciano Floridi


$\langle \chi, \pi \rangle \rightsquigarrow \langle \alpha, \mathit{argument/proof} \rangle$
?
?
h
?

Contrarian view on animal minds in *Nat. Geo.*:
http://ngm.nationalgeographic.com/2008/03/animal-minds/virginia-morell-text


$\langle \chi, \pi \rangle \rightsquigarrow \langle \alpha, argument/proof \rangle$
?
?
h
?

Contrarian view on animal minds in *Nat. Geo.*:
http://ngm.nationalgeographic.com/2008/03/animal-minds/virginia-morell-text



http://kryten.mm.rpi.edu/SBringsjord_etal_self-con_robots_kg4_0601151615NY.pdf

https://www.youtube.com/watch?v=MceJYhVD_xY

Floridi's Continuum (augmented), and Claims

	False Belief Task	Wise Man Test (*n*)	Deafening Test	Torture Boots Test	Ultimate Sifter	*Infinitary* False Belief Task
Cutting-Edge AI	Yes	Yes	No	No	No	?
Zombies	Yes	Yes	Yes	Yes	No	?
Human Persons (s-conscious! p-conscious!)	Yes	Yes	Yes	Yes	Yes	Yes

Floridi's Continuum (augmented), and Claims

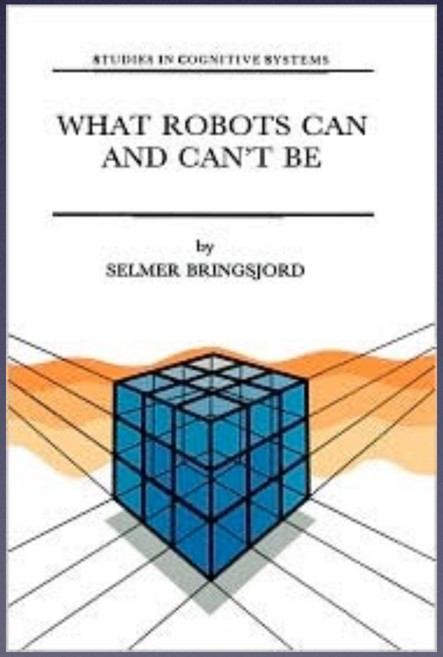
	False Belief Task	Wise Man Test (*n*)	Deafening Test	Torture Boots Test	Ultimate Sifter	*Infinitary* False Belief Task
Cutting-Edge AI	Yes	Yes	No	No	No	?
Zombies	Yes	Yes	Yes	Yes	No	?
Human Persons (s-conscious! p-conscious!)	Yes	Yes	Yes	Yes	Yes	Yes

Infinitary False Belief Task

http://kryten.mm.rpi.edu/PRES/COGSCI2019/infinitaryfalsebeliefprezCogSci2019.key

So ... despite the fact we can’t build rational persons, apparently we can build AIs that pass *any* short test. That’s why *Blade Runner* (& *Ex Machina*?) is our future.

So ... despite the fact we can’t build rational persons, apparently we can build AIs that pass *any* short test. That’s why *Blade Runner* (& *Ex Machina*?) is our future.



STUDIES IN COGNITIVE SYSTEMS

WHAT ROBOTS CAN
AND CAN’T BE

by
SELMER BRINGSJORD