Diachronic Dutch Book Arguments for Forgetful Agents

Alistair M. C. Isaac
Department of Philosophy
University of Michigan

June 29, 2011

Outline of the Talk

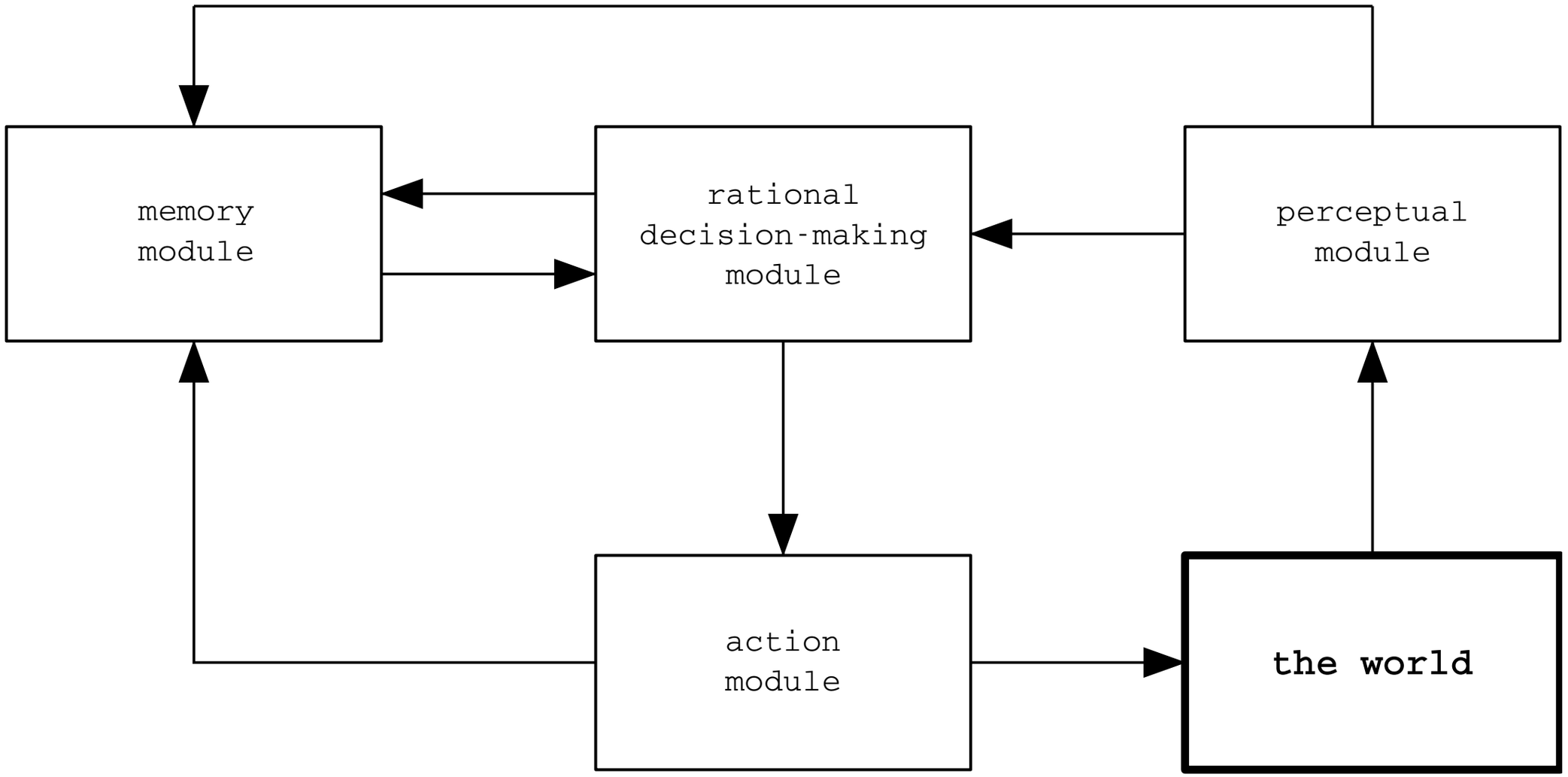
1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
    a. What are Dutch book arguments?
    b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
    a. Decision problems
    b. Imperfect recall
    c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
    a. The challenge: *Action and belief come apart*
    b. The absentminded driver
    c. Sleeping beauty

Outline of the Talk

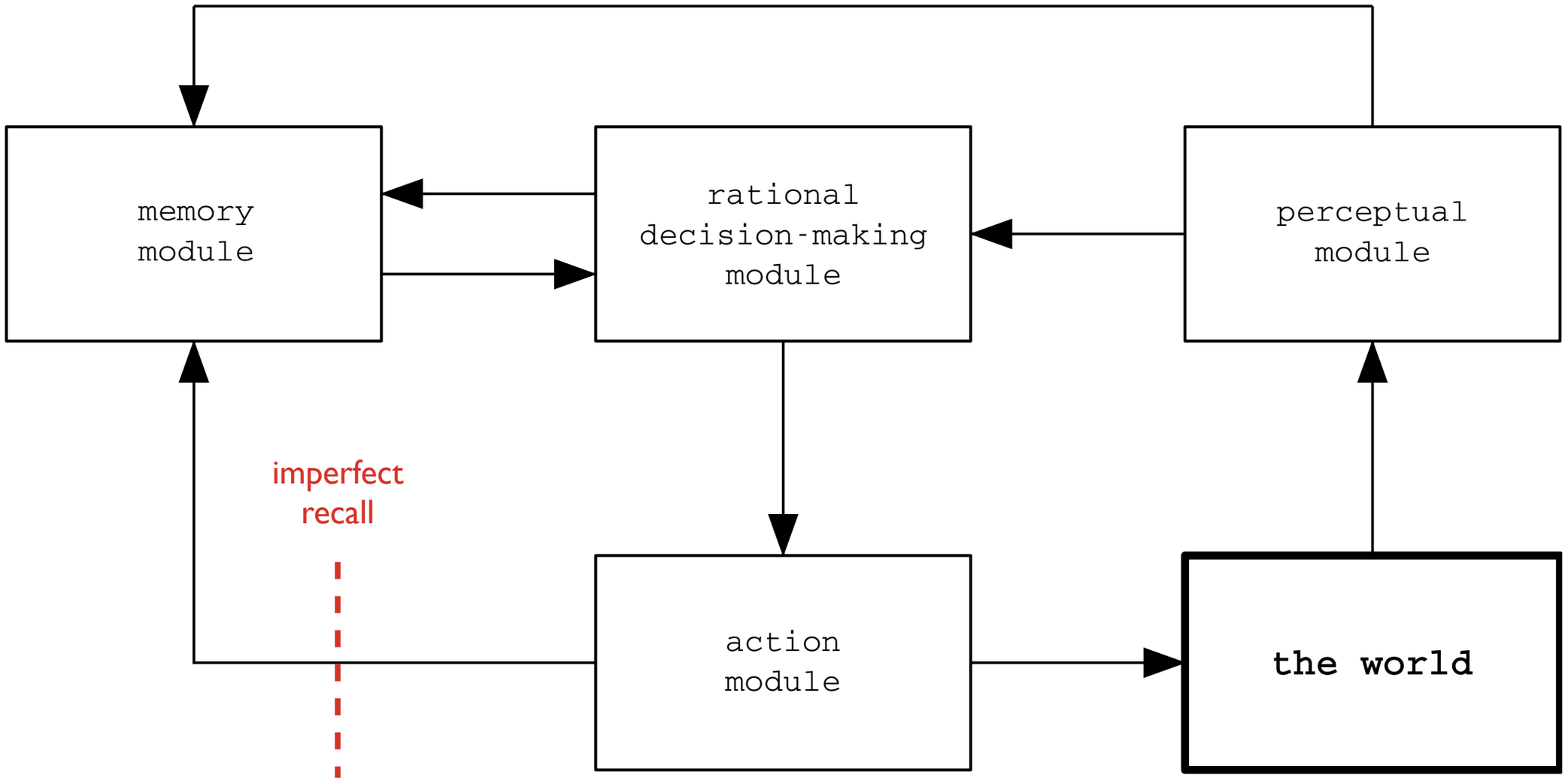
1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
    a. What are Dutch book arguments?
    b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
    a. Decision problems
    b. Imperfect recall
    c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
    a. The challenge: *Action and belief come apart*
    b. The absentminded driver
    c. Sleeping beauty

Bounded Rationality

- *Bounded rationality* - Herbert Simon introduced this term to characterize the meaning of "rationality" in a situation where the agent's computational resources are insufficient to model the full complexity of his environment.
- Simon argued that agents in such a situation would need to use heuristics, or "rule of thumb" strategies involving locally available information, in order to generate rational behavior in a complex environment.
- More generally, any agent whose informational access to the world is somehow imperfect can be thought of as a "bounded agent" . . .
- We can ask: *Given that the agent is bounded in such-and-such a way, what is the rational action for the agent to perform in a particular context?*
- (This assumes the agent is aware of his own impediment.)



- Here is a rough schematic of the information flow between world and decision-making module for a simple agent.
- By perturbing the flow of information, we can investigate rational behavior if the agent is *bounded*.



- For example, if an agent forgets an action he has performed in the past, we might perturb the relation between the action module and the memory module.
- The important point here is just that the agent's rationality is unchanged, all that has changed is his access to information about the world.

Outline of the Talk

1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
    a. What are Dutch book arguments?
    b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
    a. Decision problems
    b. Imperfect recall
    c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
    a. The challenge: *Action and belief come apart*
    b. The absentminded driver
    c. Sleeping beauty

Dutch book arguments

- A *Dutch book* is a system of bets placed with a bookie which guarantees that the bookie will always lose (and the bettor will always win).
- A *Dutch book argument* demonstrates that an agent is susceptible to a Dutch book if he does not follow the recommendation of a particular epistemic norm.
- Dutch book arguments in this sense were proposed by Frank Ramsey (1926). His essential insight was that the nebulous quantity "degree of belief" is only interesting *as a guide to action*. As such, we can *operationalize* and *measure* degrees of belief by considering the particular action of betting.
- If an agent assigns *inconsistent* degrees of belief, a cunning bettor can take advantage of that inconsistency to place bets with him such that he always loses. This is the appeal of Dutch book arguments in the subjectivist context: *they ensure the* **consistency** *of degrees of beliefs*.

Dutch book arguments

- Dutch book arguments have been used by subjectivists about probability to demonstrate several norms:
  - That assignment of degrees of partial belief should satisfy the axioms of probability (de Finetti, 1937)
  - $P(A\&B) = P(A|B)P(B)$ (de Finetti, 1937)
  - Agents should update by *conditionalization*, i.e. if A is observed between time 1 and time 2, for all $X \in \Omega$, $P_2(X) = P_1(X|A)$ (Lewis / Teller, 1973)
- Notice that the last of these norms is qualitatively different from the first two:
- The results proved by Bruno de Finetti place constraints on an agent's *synchronic* belief state, how that agent assigns (conditional) degrees of belief *at a particular time*.
- The result of David Lewis (reported by Paul Teller), however, places a *diachronic* constraint on beliefs: it constrains the change in beliefs over time.

Outline of the Talk

1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
    a. What are Dutch book arguments?
    b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
    a. Decision problems
    b. Imperfect recall
    c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
    a. The challenge: *Action and belief come apart*
    b. The absentminded driver
    c. Sleeping beauty

Two Diachronic Worries

- *Diachronic* Dutch book arguments involve systems of bets placed at multiple points in time.
- This feature has opened the door for criticisms which do not apply to synchronic Dutch book arguments.
- Objections to diachronic Dutch book arguments take two general forms:

1. Appeal to strategic features of the betting scenario, or
2. Rejection of consistency as a constraint on beliefs across time.

Strategic Objections

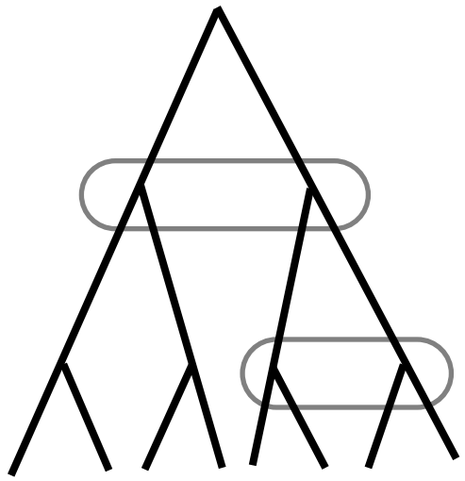
- Appeals to strategic features of a diachronic betting scenario are those which appeal to new actions available to the agent (e.g. not accepting a bet) or informational changes predicated on interpreting the bet as a strategic interaction between different agents (e.g. "seeing the Dutch book coming").
- Such objections can be found in Maher (1992), Border and Segal (1994), and many other authors. Bradley and Leitgeb (2006), for example, compare bets offered to forgetful agents with bets offered to hallucinating agents, or bets involving forged money.
- From the subjectivist standpoint, however, these strategic moves simply block a test of the *consistency* of the agent's beliefs.
- To refuse to accept a bet in this context means *opting out of the test of consistency*. This is why it is helpful to think of the agent as a bookie - *bookies offer odds*, it is their job to accept bets on those odds on outcomes and monetary amounts determined by the bettor. *To refuse such a bet is to stop being a bookie*.

Strategic Objections

- In order to understand a Dutch book argument *as a test of consistency*, it is perhaps better to think of the bets as *a kind of thought experiment* the agent makes with himself, or a bet between two agents *with exactly the same informational access to the world*.
- If the bettor knows more than the bookie (e.g. that a certain outcome has already obtained, that one bet will be placed more frequently than another, that some bets are made with forged money, etc.), then *of course* the bettor can Dutch book the bookie (this is in fact how Dutch Schultz implemented his Dutch books) - but this is no test of the consistency of the bookie's beliefs (*nor of his rationality!*).
- In the context of bounded agents, then, we must ensure that *the bettor is bounded in exactly the same way as the bookie* in order for the diachronic Dutch argument to be legitimate *as a test of consistency*.

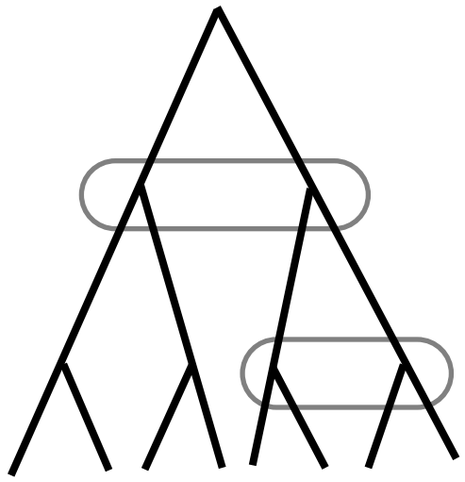
Consistency Objections

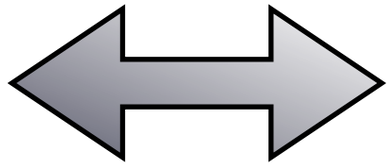
- Even if this response succeeds, and we acknowledge that strategic considerations are irrelevant to tests of consistency, a second type of objection still remains: *What if consistency is an illegitimate norm to apply to beliefs across time?*
- Suppose it is 6 pm now and I know that at 10 pm I will be drunk. Suppose further I believe now that I will be unable to drive at 10 pm, but I also know that *in virtue of my impaired state*, I will believe at 10 pm that I am able to drive. *Surely it would be* ***irrational*** *if I adjusted my beliefs now to be consistent with my beliefs at 10 pm!*
- Similar considerations apply to cases involving forgetfulness, hallucinations, or any situation involving a superior epistemic state now and a future impaired state.
- *The purpose of the rest of this talk is to respond to this type of objection.*



DDB theorems
are proved on formal
models, which may
involve bounded agents

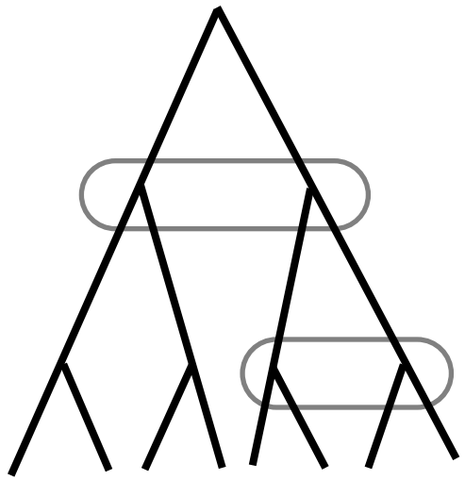
- The basic idea is to show how diachronic Dutch book theorems can be proved for precisely specified decision problems involving forgetfulness.
- Decision problems are just ways in which the world can evolve and events can occur, something like extensive form games, or the tree models commonly used for dynamic / temporal epistemic logics.



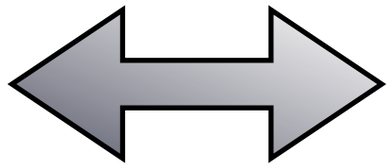


DDB theorems
are proved on formal
models, which may
involve bounded agents

- A separate argument is needed to show that these models correspond to decision making problems of interest in the real world.
- If we start from a problem about the real world, we need to make sure we've included all its relevant features in a formal model before any results we prove about that model (e.g. a DDB theorem) can be said to be relevant to the real world.

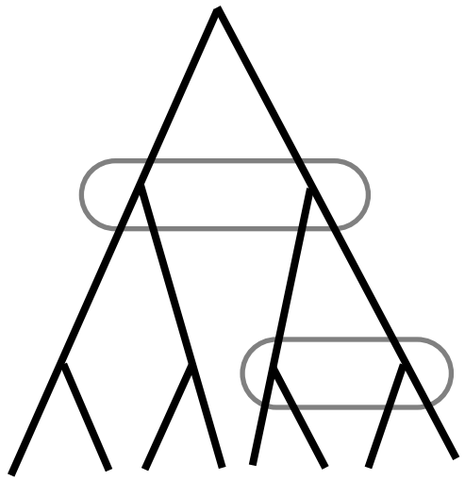


DDB theorems
are proved on formal
models, which may
involve bounded agents

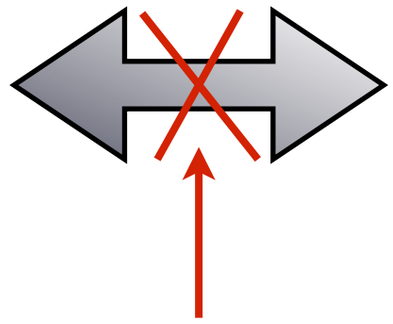


. . . but these models
may not capture all
the relevant details of
the real world decision
problem of interest.

- I think the worries that consistency is not an appropriate norm for forgetful (more generally, bounded) agents, are not problems for DDB arguments about forgetful agents, per se . . .
- . . . but rather worries about the closeness of fit between the models used to prove such results and the real problem of interest.
- My response: *the burden falls on you to identify all relevant features of the problem so we can formalize it and investigate its properties!*



DDB theorems
are proved on formal
models, which may
involve bounded agents



This is the
real locus of
"consistency"
worries!

. . . but these models
may not capture all
the relevant details of
the real world decision
problem of interest.

- I think the worries that consistency is not an appropriate norm for forgetful (more generally, bounded) agents, are not problems for DDB arguments about forgetful agents, per se . . .
- . . . but rather worries about the closeness of fit between the models used to prove such results and the real problem of interest.
- My response: *the burden falls on you to identify all relevant features of the problem so we can formalize it and investigate its properties!*

Outline of the Talk

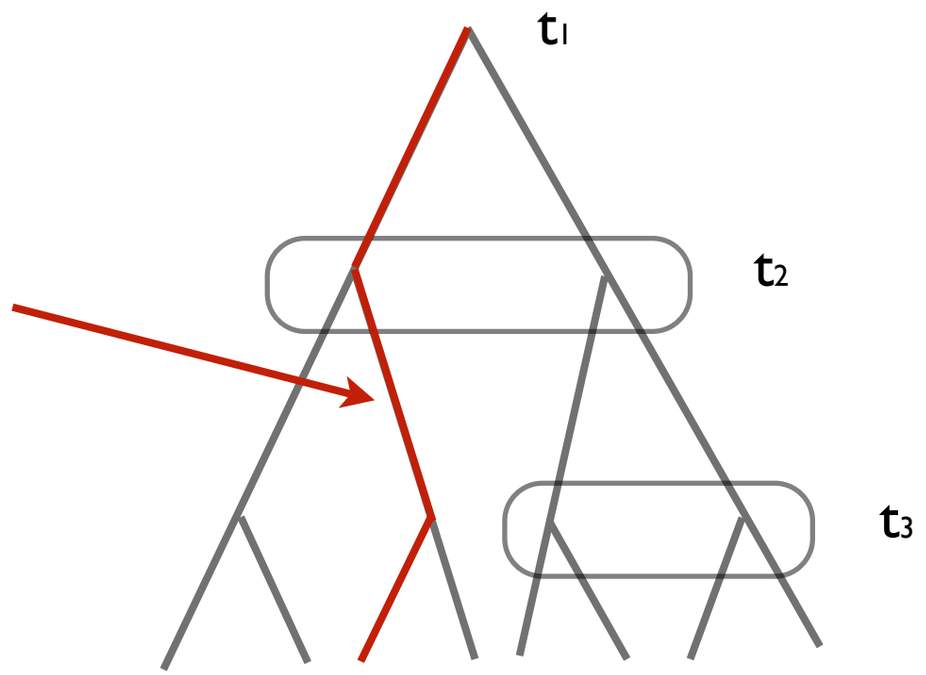
1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
   a. What are Dutch book arguments?
   b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
   a. Decision problems
   b. Imperfect recall
   c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
   a. The challenge: *Action and belief come apart*
   b. The absentminded driver
   c. Sleeping beauty

Decision Problems

- In order to prove a diachronic Dutch book argument, we need to specify the properties of the agent of interest and a *decision problem*.
- I treat decision problems as a sequence of time steps $t_1, t_2, t_3, \ldots$ between each of which some information is presented to the agent: between t_1 and t_2 he might learn that A is true, between t_2 and t_3 he might learn that P(B) = .8, etc.
- The agent updates his belief state at each time step, and belief states are represented by probability distributions.
- So, we may ask, given the agent's belief state at t_1 is P_1, and he learns between t_1 and t_2 that A is true, how should P_2 be defined?
- In this case, Lewis' argument shows that for all $X \in \Omega$, $P_2(X) = P_1(X|A)$.
- An intuitive way to think of decision problems is as branching tree structures. Ω, the set over which beliefs are defined, is the set of histories (or paths) through the tree.

Trees

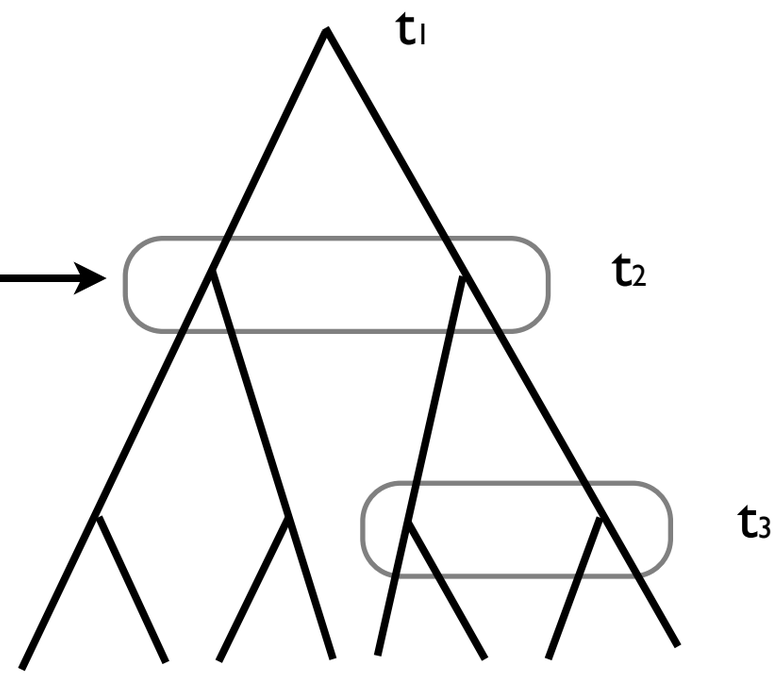
A history through the tree. Ω is the set of all histories.



- Ω is the set of all histories in the tree. Intuitively, this is the set of different ways the world could be. *Probabilities are defined on an algebra over* Ω, though for simplicity's sake, I will sometimes just refer to this algebra as Ω.
- [Note that we do not consider probabilities over probabilities (e.g. claims that P(*ch*(A) = .8) = 1, where *ch* is an "objective probability function") - the relevant probability space will need to be defined explicitly in order to apply the methods discussed here.]

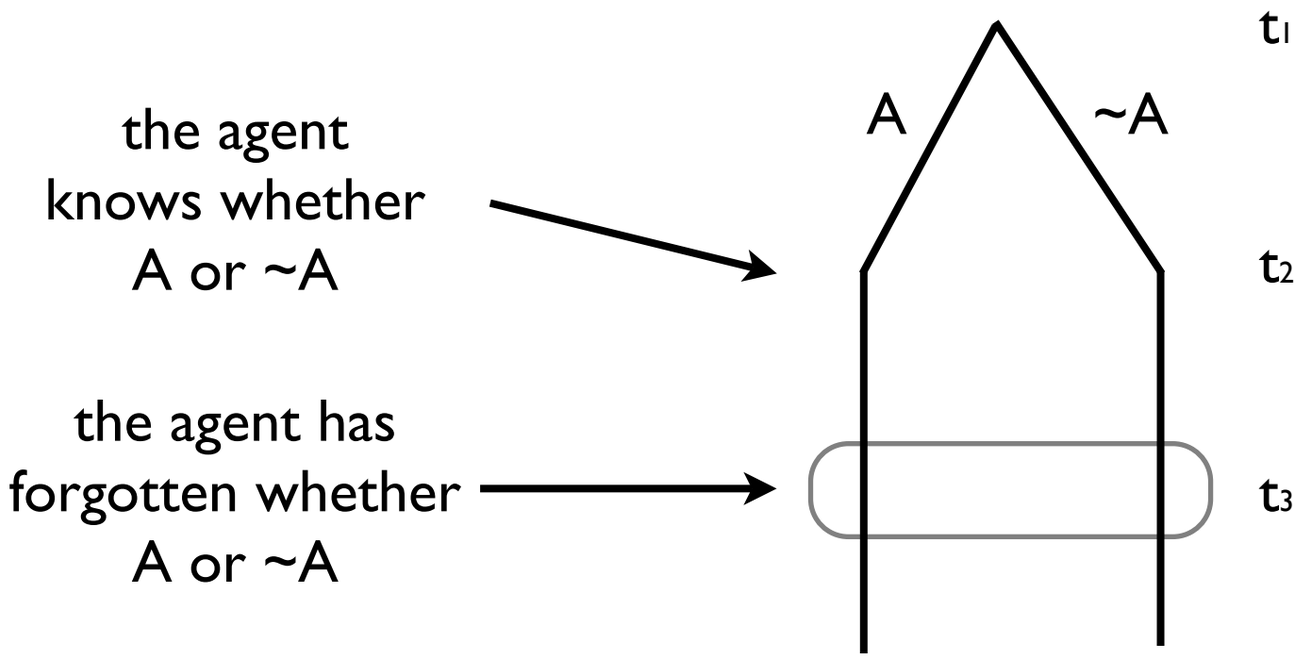
Trees

These nodes are indistinguishable for the agent



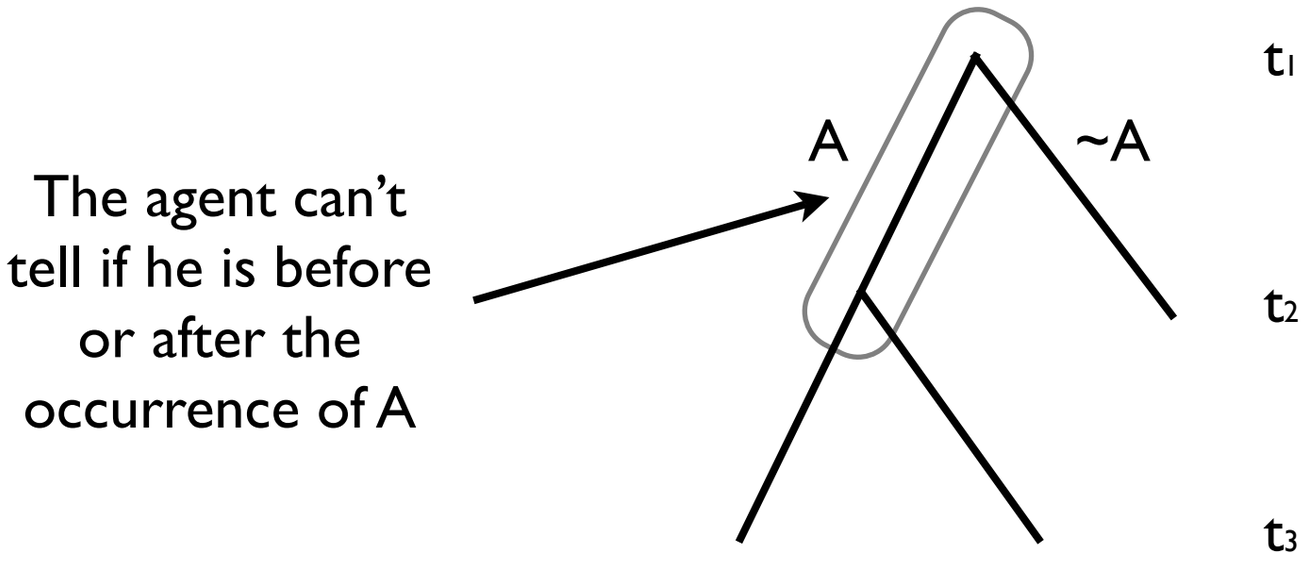
- When we are considering bounded agents, we can think of some of the paths through the tree as being indistinguishable for the agent.
- Trees with indistinguishability relations defined on them are used in game theory (extensive form games) and modal logic (e.g. DEL, ETL).
- I will use some terminology from game theory to characterize different types of forgetfulness.

Imperfect Recall



- In game theory, if an agent performs an action, then later forgets which action he performed, he experiences **imperfect recall**.
- More generally, if an agent knows something (whether from performing the action himself, or making a veridical observation, whatever), then later forgets it, we say he has imperfect recall.
- In this example, the agent observes whether A or ~A occurs between t_1 and t_2, but between t_2 and t_3, the agent forgets which of A or ~A occurred.

Absentmindedness



- In game theory, if a set of indistinguishable nodes includes two nodes which fall on the same history, the agent is **absentminded**.
- An absentminded agent is worse off than one with imperfect recall - he not only doesn't know what event occurred, he doesn't know whether it is before or after the occurrence of the event.
- In the above example, the agent can't tell whether it is t_1 and it is still indeterminate whether A or ~A, or whether it is t_2 and A has occurred.

Outline of the Talk

1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
    a. What are Dutch book arguments?
    b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
    a. Decision problems
    b. Imperfect recall
    c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
    a. The challenge: *Action and belief come apart*
    b. The absentminded driver
    c. Sleeping beauty

Diachronic Dutch Books

- An agent in a decision problem updates his probability distribution in accordance with a function UPDATE from his initial probability distribution and his evidence to a new probability distribution.
- For example, in Lewis' argument, UPDATE: $P_t \times \Omega \rightarrow P_{t+1}$
- A Dutch book argument shows how a Dutch book can be constructed from UPDATE if it fails to satisfy the relevant norm; in the Lewis case, if UPDATE fails to conditionalize on the evidence, which we represent by E.
- Suppose, for example, $P_2(\cdot) = \text{UPDATE}(P_1, E) \neq P_1(\cdot|E)$, then for some A in Ω, either $P_2(A) > P_1(A|E)$ or $P_2(A) < P_1(A|E)$. Consider the second case, and set $\varepsilon = P_1(A|E) - P_2(A)$.
- At t_1, the bettor places two bets with the bookie, a bet conditional on E to win \$1 if ~A at the bookie's fair price of $\$P_1(\sim A|E)$ and a side bet on ~E to win $\$\varepsilon$ at the bookie's fair price of $\$\varepsilon P_1(\sim E)$.
- At t_2, if ~E has occurred, the first bet is not placed, and the bettor gets a net with of $\$(1 - P_1(\sim E)) = \$\varepsilon P_1(E)$.

Diachronic Dutch Books

- At t_2, if E has occurred, the first bet is placed, and the bettor places a third bet with the bookie to win $1 if A at the bookie's fair price of $\$P_2(A)$.
- Now the bettor's net win is always positive:
- if ~E, $\varepsilon P_1(E)$ by bet 2
- if E&A, $P_2(\sim A)$ [bet 3] - $P_1(\sim A|E)$ [bet 1] - $\varepsilon P_1(\sim E)$ [bet 2] = $\varepsilon P_1(E)$
- if E&~A, $P_1(A|E)$ [bet 1] - $P_2(A)$ [bet 3] - $\varepsilon P_1(\sim E)$ [bet 2] = $\varepsilon P_1(E)$
- An analogous set of bets can be constructed for the case $P_2(A) > P_1(A|E)$.
- So, from a discrepancy between what one accepts conditionally at one time, and one's updated belief at a later time, a Dutch book can be constructed. So, the only update strategy which preserves consistency between these times is conditionalization.

The Skyrms Strategy

- Brian Skyrms (1987) proved a diachronic Dutch book argument for Jeffrey conditionalization.
- Jeffrey conditionalization applies when one receives evidence which reapportions probabilities over a partition of Ω. Call the partition $\{E_i\}$ and the new weights on each E_i, q_i (with $\sum q_i = 1$), then Skyrms' argument concerns a decision problem such that

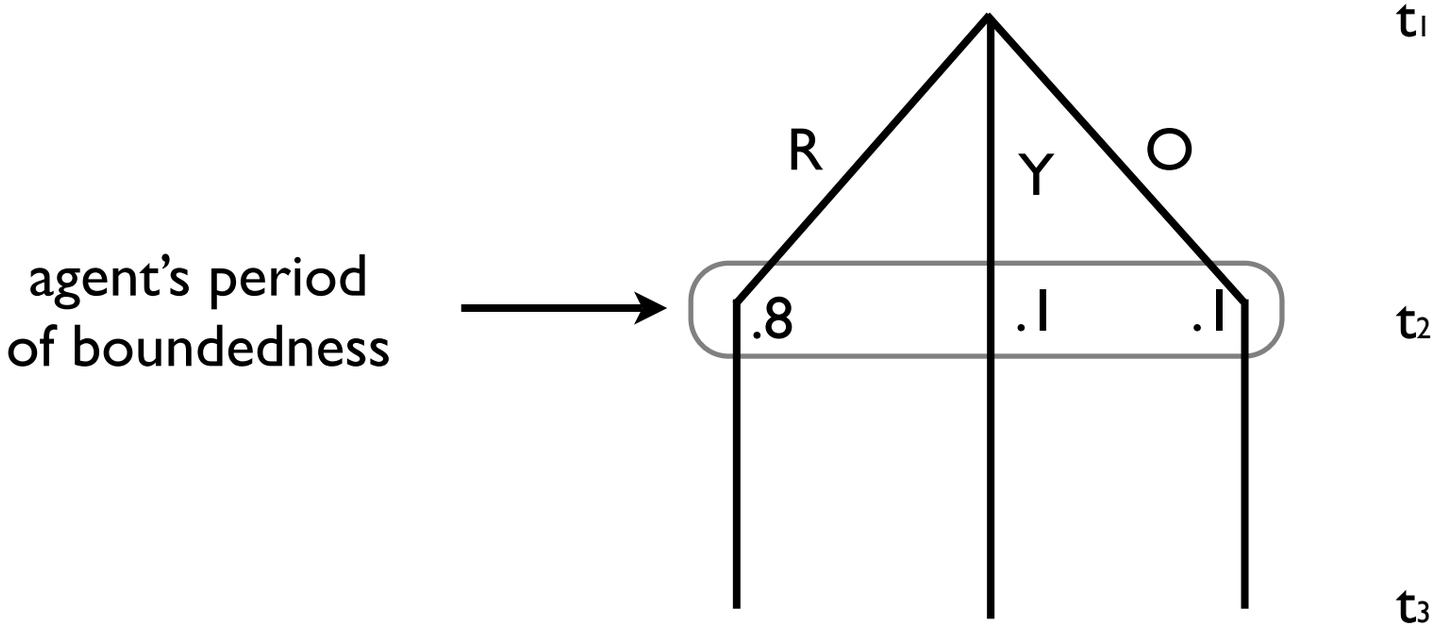
  1. There are three time steps, and between t_1 and t_2, the agent learns a weighting $\{q_i\}$ on the partition $\{E_i\}$. Between t_2 and t_3, the agent learns that for some a, E_a is true (equivalently, $q_a = 1$).
  2. The agent updates by a function UPDATE: $P_t \times [\{E_i\}, \{q_i\}] \rightarrow P_{t+1}$

- Jeffrey conditionalization ensures that all probabilities conditional on members of $\{E_i\}$ remain unchanged. Given P and $[\{E_i\}, \{q_i\}]$, it first sets $P^*(E_i) = q_i$, then for all A in Ω, it sets $P^*(A) = \sum P(A|\ E_i)P^*(E_i)$.
- Skyrms shows UPDATE must be equivalent to Jeffrey's rule, or the agent can be Dutch booked.

The Skyrms Strategy

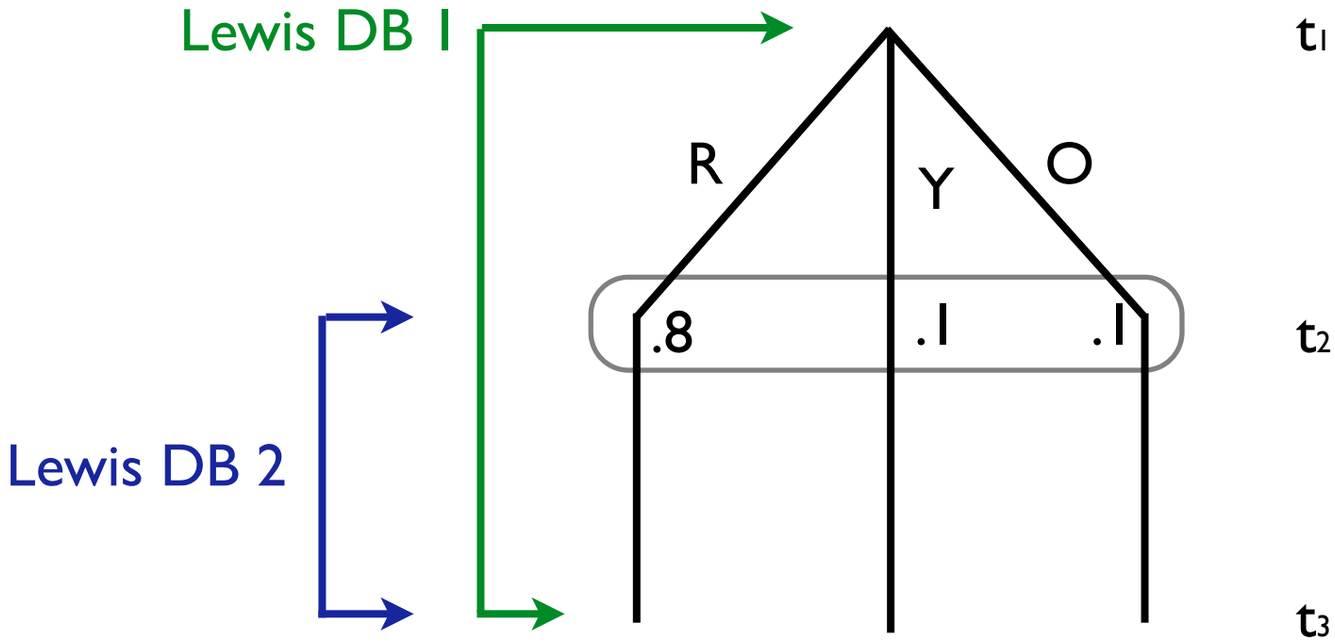
- An intuitive problem which may employ Jeffrey conditionalization is indistinct perception. For example, at t_1 you know your wife just changed into a new dress, and you have some arbitrary probability distribution over colors.
- Between t_1 and t_2, you see your wife in her new dress in a poorly lit room, maybe against a flickering candle. This gives you some evidence, but not enough to uniquely determine the color of the dress, maybe you now assess the probability the dress is red at .8, the probability it is yellow at .1, and the probability it is orange at .1.
- Then at t_3, you walk out into the sunlight and observe definitively that the dress is orange.
- Here, the partition is {*red*, *yellow*, *orange*}, q_1=.8, and $q_2 = q_3 = .1$.

The Skyrms Strategy



- In order to probe the rational assignment of belief during a period of boundedness (i.e. after the dress has been observed indistinctly by candlelight), Skyrms considers the times which *bracket* that period, t_1 and t_3, which are both *epistemically secure*.
- Call a node **epistemically secure** if the agent knows which past history brought him to that node. Call a decision problem **bracketed** if both the start and end nodes are epistemically secure.

The Skyrms Strategy



- By considering a bracketed decision problem, Skyrms can simply apply multiple Lewis-style Dutch books to get his result.
- This ensures that conditional probabilities stay stable during the decision problem, and thus that updating must have occurred at both steps via Jeffrey conditionalization.
- **Research program**: *Bracket other types of bounded decision problem and apply Skyrms' strategy to prove diachronic Dutch books.*

Outline of the Talk

1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
    a. What are Dutch book arguments?
    b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
    a. Decision problems
    b. Imperfect recall
    c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
    a. The challenge: *Action and belief come apart*
    b. The absentminded driver
    c. Sleeping beauty

The Spaghetti Dinner

- Talbott (1991) posed the following problem for diachronic Dutch book arguments:
- On June 29, 2010, Talbott eats spaghetti for dinner. One year later, on June 29, 2011, Talbott is asked to offer odds on what he ate for dinner one year previously. Talbott doesn't remember exactly what he had for dinner, all he remembers is that one year ago he had three basic types of dinner, spaghetti, roast beef, and tacos, each of which he ate about a third of the time. So, Talbott offers odds of 1/3 on spaghetti for dinner a year previously.
- Talbott's claim: since he would offer odds of spaghetti for dinner on the evening of June 29, 2010 at $P_{2010}(S) = 1$, but one year later, he would offer odds $P_{2011}(S) = 1/3$, he can be Dutch booked.
- Nevertheless, he does not seem to act *irrationally* here by forgetting something as unimportant as what he had for dinner on June 29, 2010.

The Spaghetti Dinner

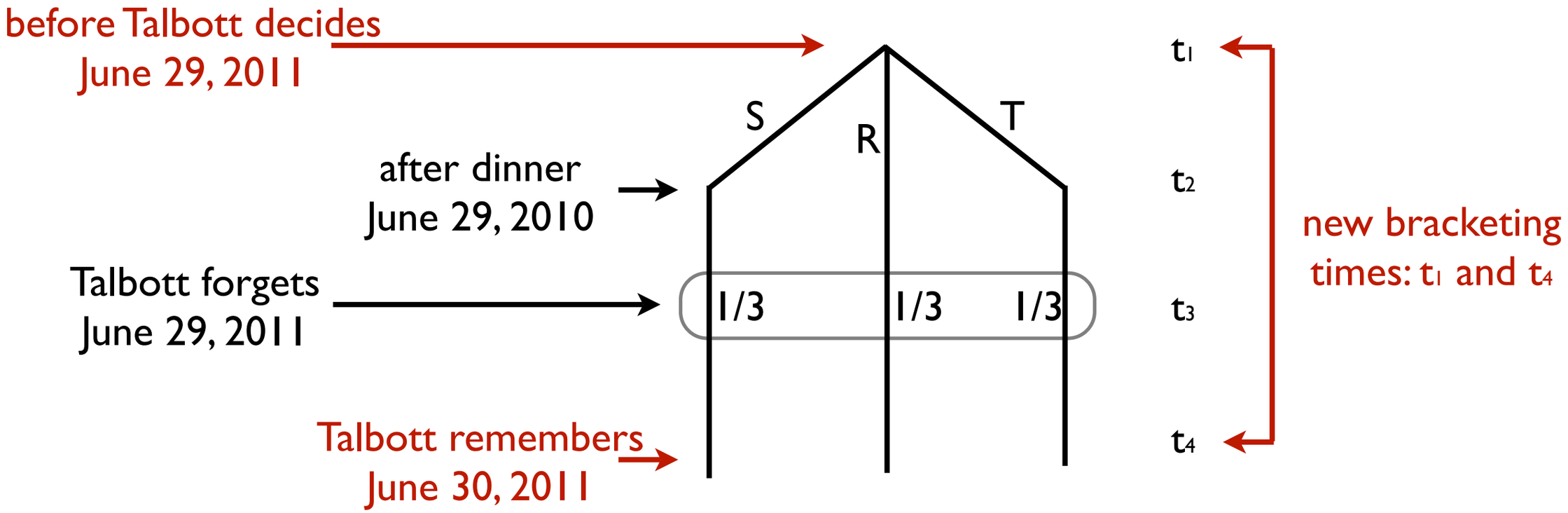
- By the lights of the present project, Talbott does not provide a legitimate Dutch book *argument* (i.e. his susceptibility to being Dutch booked here is not evidence of irrationality) for two reasons:

  1. Talbott's Dutch book depends upon illegitimate strategic considerations: the bettor must remember what bet was placed a year ago in order to ensure a Dutch book. If this information is available to the bettor, but not Talbott, then the bettor's ability to Dutch book Talbott does not say anything about the consistency of Talbott's beliefs, nor about his rationality.

  2. This decision problem is not bracketed, in particular, we are testing the consistency of Talbott's beliefs during a period of boundedness (in this case, imperfect recall), but only one end of the decision problem (the first) is epistemically stable.

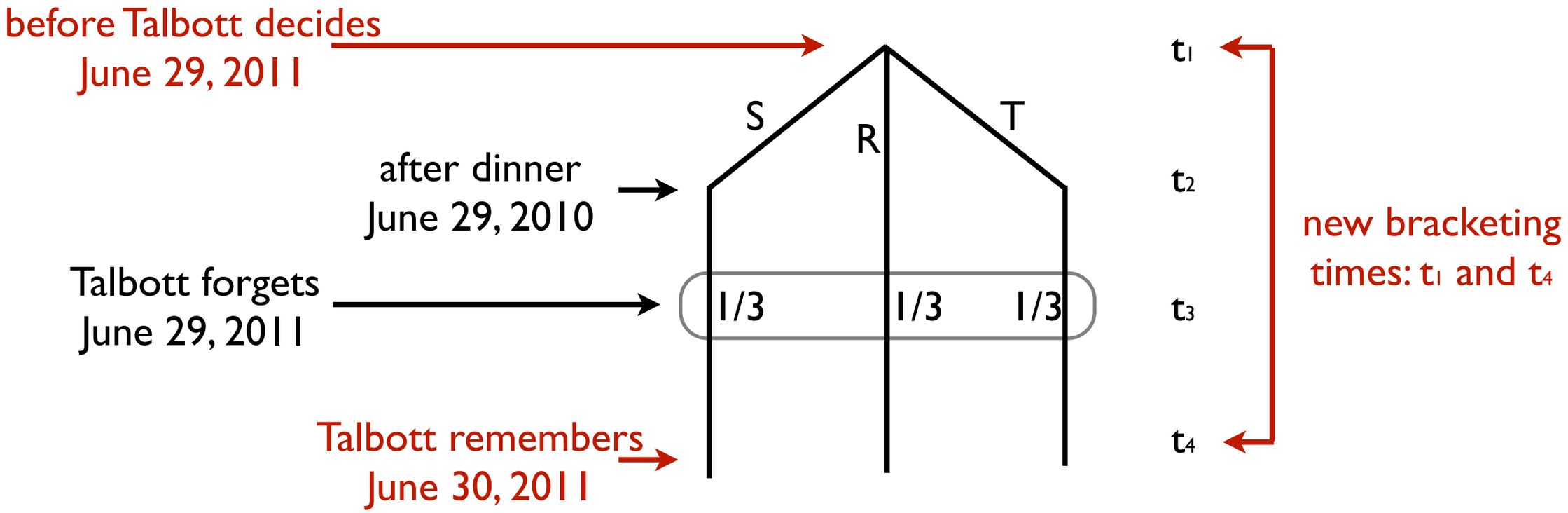
- In fact, in order to prove a Skyrms-style Dutch book theorem about the spaghetti dinner, we will need an even wider bracket, one which considers an earlier node than that at which Talbott's bettor places his first bet.

The Spaghetti Dinner



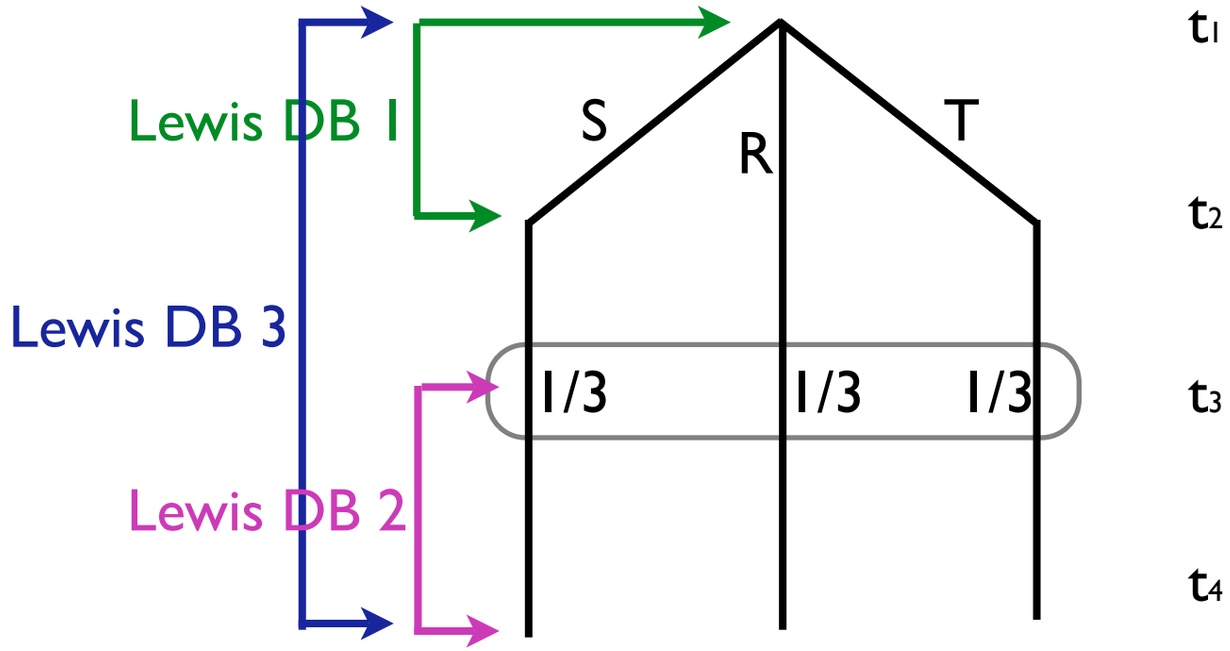
- To Talbott's decision problem, which involves two times, after dinner in 2010 and after forgetting in 2011, we add two new times: an earlier time, before he has decided on what to have for dinner in 2010; a later time, after his period of forgetfulness, during which he remembers what he had for dinner in 2010 (perhaps via a memory aid, e.g. he stumbles across an old diary entry which describes his dinner of June 29, 2010).

The Spaghetti Dinner



- The earlier time allows us to check the consistency of Talbott's beliefs during his forgetfulness. They will not be consistent with his beliefs at t_2, since he will now assign values to propositions formerly assigned zero. e.g. $P_2(R) = 0$, but $P_3(R) = 1/3$.
- Considering a time after Talbott has remembered will help us check that his beliefs remain consistent if he exits his period of boundedness.

The Spaghetti Dinner



- We can now construct a Skyrms-style Dutch book argument for the Spaghetti dinner, demonstrating (for whatever $X \in \{S, R, T\}$ is true):

1. $P_2(\cdot) = P_1(\cdot|X)$
2. $P_3(\cdot) = \sum P_1(\cdot|Y)P^*(Y)$, where $P^*(Y) = 1/3$ for each $Y \in \{S, R, T\}$
3. $P_4(\cdot) = P_3(\cdot|X)$ [$= P_1(\cdot|X) = P_2(\cdot)$]

The Spaghetti Dinner

- What is the significance of this result?

1. This strategy should generalize to other simple examples of *imperfect recall* (as long as the relevant new bracketing times are available).

2. Should it matter that we had to add a time when Talbott remembers to the problem? Maybe not: surely you'd want to update belief consistent with the *possibility* of remembering. Furthermore, "remember" here just means "receives veridical information that," which taken generally seems more plausible (e.g. finding an old diary entry).

3. Notice: we upheld Talbott's conclusion (if not his argument) - we do not condemn the forgetful agent for an inconsistency between t_2 and t_3. Rather, we take *as given* his forgetfulness at t_3, and ask what demands should we make of the rest of his beliefs in order to maintain diachronic consistency?

Real Life Spaghetti Dinners

- So, if you are convinced by the preceding, you should accept that we have a strategy for checking the consistency of belief states given imperfect recall *in formally specific decision problems*.
- *Do these results extend to real world examples?* In order to be rational, must my belief state today have been derived by applying Jeffrey conditionalization to my state of belief before I decided on dinner June 29, 2010?
- Obvious answer is NO. The reason, however, is that *many intermediary actions have been taken and changes of belief state occurred* since 2010.
- These intermediary actions and informational changes are not included in our formalization of the spaghetti dinner problem. So, the DB theorem applies only to real life decision problems where the **only** choice points of interest correspond to nodes in the formalization.
- In the case of imperfect recall (normal forgetfulness), *the relevant nodes are usually very far apart*, and correspondingly have many choice points of interest between them - so it seems that relatively few decision problems will correspond to this one in the real world.

Morals on Imperfect Recall

- We can prove diachronic Dutch book theorems about simple decision problems with imperfect recall.
- In order for these results to be relevant to real world forgetfulness, all relevant changes of information must be included in the formal model.
- In general, the farther apart the informational changes involving forgetfulness in the real world, the less likely this will be.

Outline of the Talk

1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
   a. What are Dutch book arguments?
   b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
   a. Decision problems
   b. Imperfect recall
   c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
   a. The challenge: *Action and belief come apart*
   b. The absentminded driver
   c. Sleeping beauty

Absentmindedness

- *Absentmindedness* occurs when an agent is uncertain of where he falls in time; in particular, he is uncertain whether a particular event has occurred (or action has been taken), or not.
- This seems like a more severe cognitive impediment than imperfect recall.
- However, those plausible examples which do exist tend to involve decision points which are quite close together in time, so *perhaps simple decision problems with absentmindedness will be more relevant to the real world.*
- I walk into the kitchen intending to turn off the stove and am distracted by a shout from the other room. I run out of the kitchen to see what's the matter. Suddenly, I realize that I am uncertain whether I turned the stove off or not.
- This is the type of problem which may be modeled as absentmindedness, and the temporal closeness of the moments of absentmindedness makes plausible the claim that there are no additional decision points of interest between them, allowing our formal model to be relevant to the world.

Absentmindedness

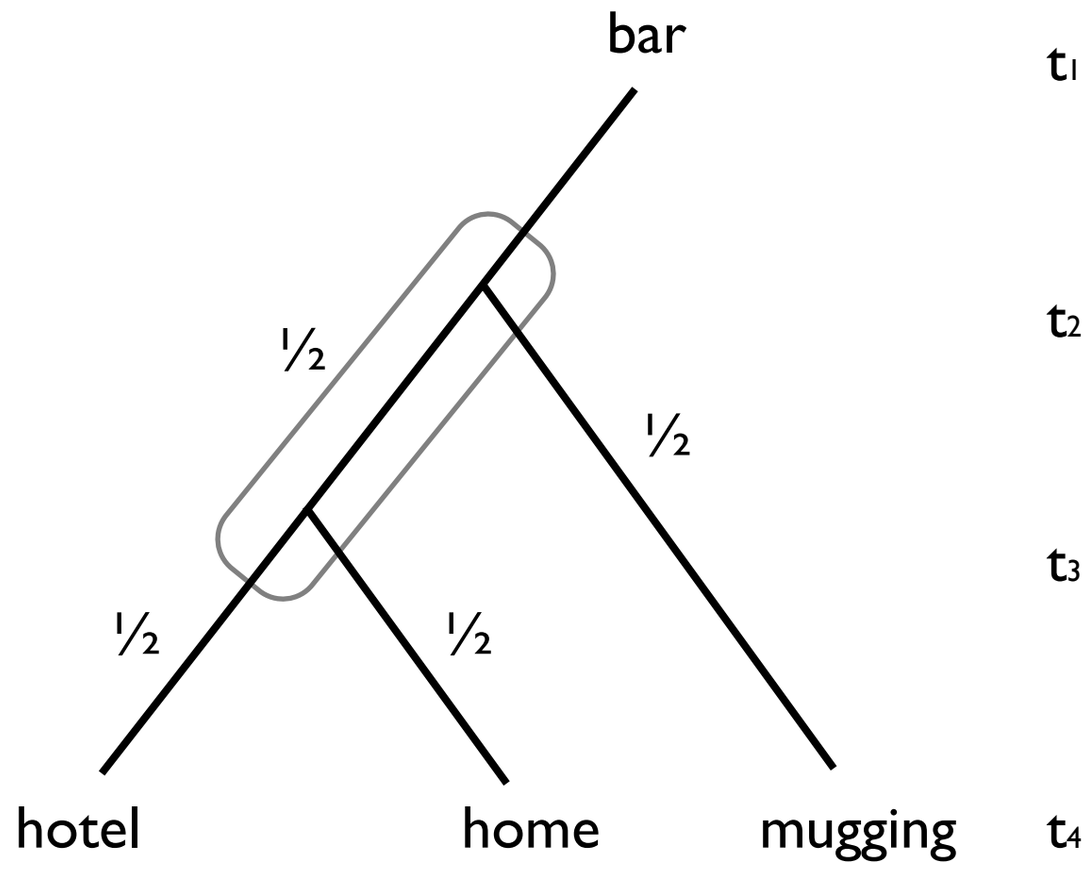
- So, decision problems with absentmindedness may fit more closely circumstances of absentmindedness in the real world, and thus formal results about them may be more relevant for real world decision making.
- However, there is an important problem for calculating correct belief in cases of absentmindedness.
- **An absentminded agent may perform an action multiple times while in a *single* belief state.**
- Thus, *actions and beliefs come apart*. (In the words of Bradley and Leitgeb (2006), "credence and betting odds diverge.")
- However, *pace* Bradley and Leitgeb, we can still *calculate* the correct (consistent) betting behavior if we add a little more detail to our description of the diachronic betting procedure. We do this by performing a kind of reverse *expected utility* analysis.

Outline of the Talk

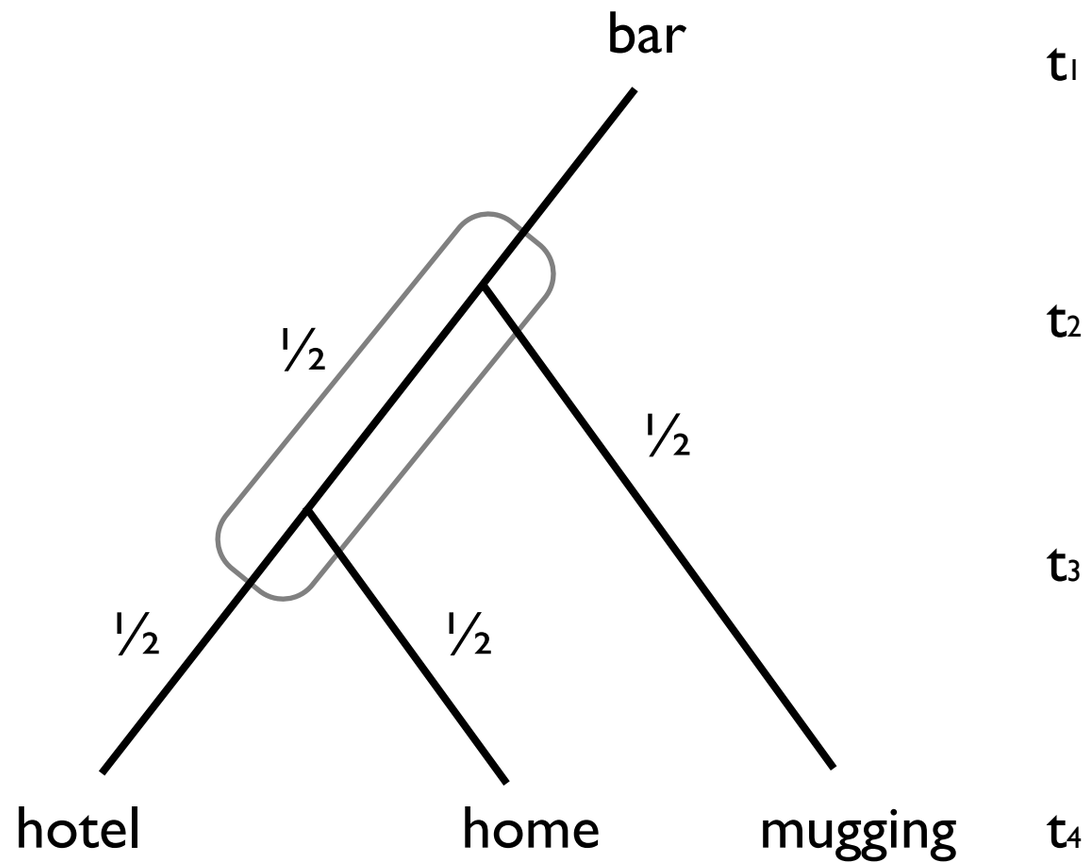
1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
   a. What are Dutch book arguments?
   b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
   a. Decision problems
   b. Imperfect recall
   c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
   a. The challenge: *Action and belief come apart*
   b. The absentminded driver
   c. Sleeping beauty

The Absentminded Driver

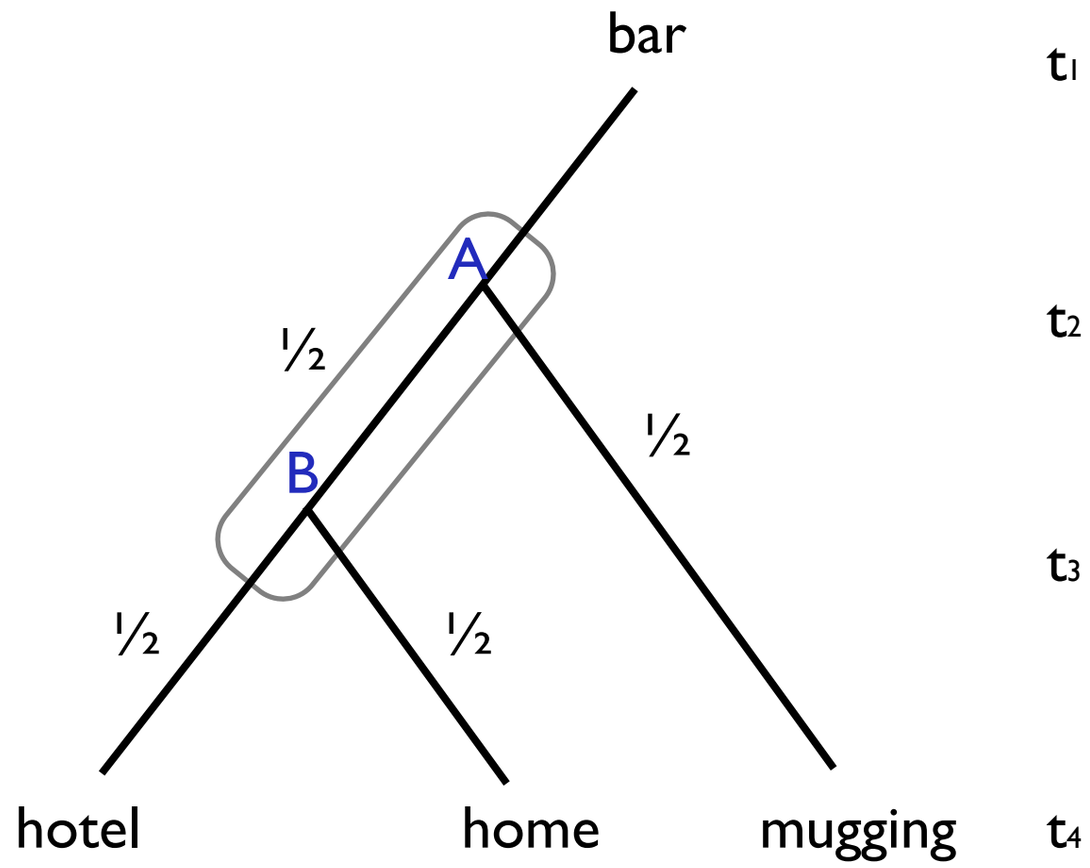
- Absentmindedness was introduced into the game theory literature by Piccione and Rubinstein (1997) with the example of the "absentminded driver." A completely passive version of this decision problem was presented by Aumann, et al. (1997) as the "forgetful passenger" - and this is the version we will discuss.
- The passenger gets drunk at a bar. He knows that there are two intersections on the highway leaving the bar which look very much the same, and in his inebriated state he will find indistinguishable. If his car turns at the first intersection, he will end up in the wrong part of town and get mugged. If his car turns at the second intersection, he will arrive safely home. If he goes straight through both intersections, he will overshoot his home and have to stay in a hotel.
- The passenger knows that his chauffeur decides where he will drive by chance. He flips two coins while waiting for the passenger to be ready to leave. If the first comes up tails, he will turn at the first intersection. If heads, he will go straight and consult the second coin flip, turning at the second intersection if tails and traveling straight if heads.



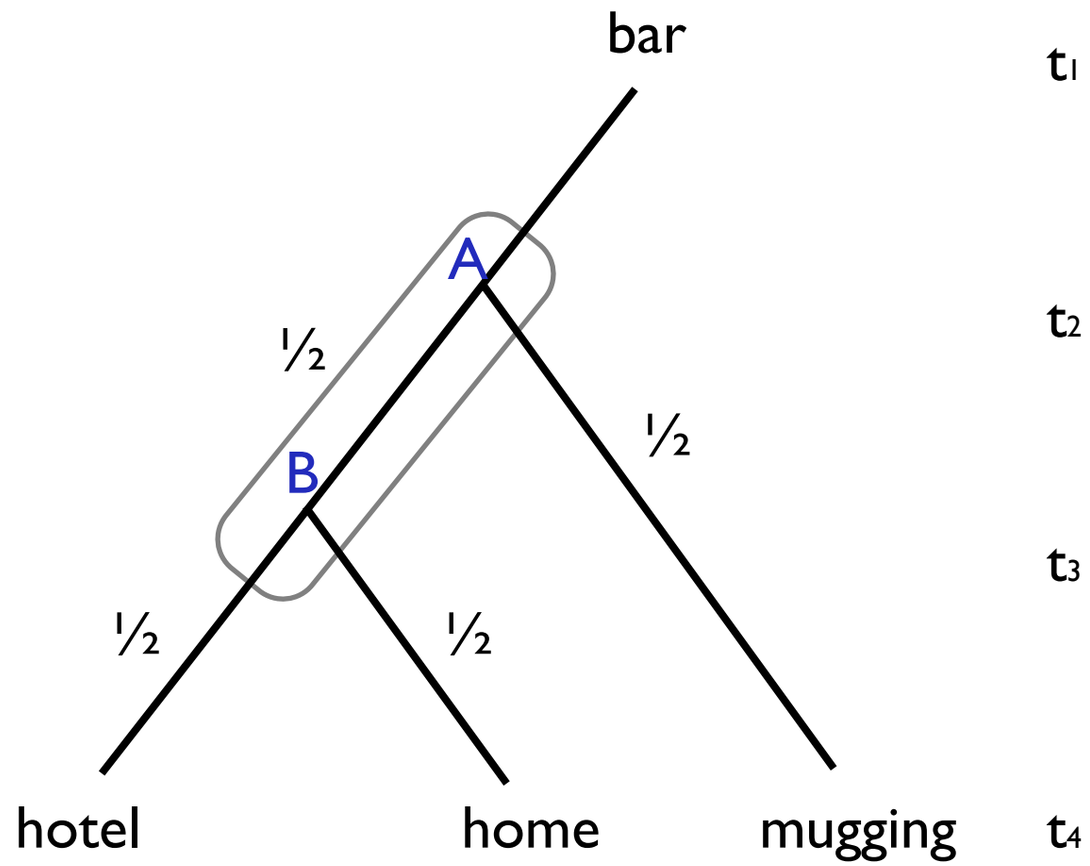
- The two intersections constitute decision points which are epistemically indistinguishable for the absentminded passenger.
- Consequently any action he rationally performs at the first, he also rationally performs at the second.



- This created a puzzle in the game theory literature - it appears, for example, as if the passenger's belief that the outcome of the first coin toss was T should be 1/2 at the bar (since he assumes the coin was fair), but 1/3 during his period of absentmindedness.
- Why should this be? We can see it by placing a Dutch book with the passenger, if he bets at t_1 as if $P_1(T) = 1/2$, then he must bet at t_2 and t_3 as if $P_2(T) = P_3(T) = 1/3$ in order to maintain consistency.



- The passenger arrives at the first intersection ("A") with probability 1, he arrives at the second ("B") with probability 1/2. During his period of absentmindedness, however, he needs to assign probability to all the nodes in his information set, so, normalizing produces
- $P_{Ab}(A) = 2/3 = P(A) / P(A) + P(B)$; $P_{Ab}(B) = 1/3 = P(B) / P(A) + P(B)$
- [Notice that these "self-locating" beliefs (P_{Ab}) are defined over the set Z of *decision nodes* across which the agent is absentminded, while *bets* are made on members of Ω, the set of histories.]



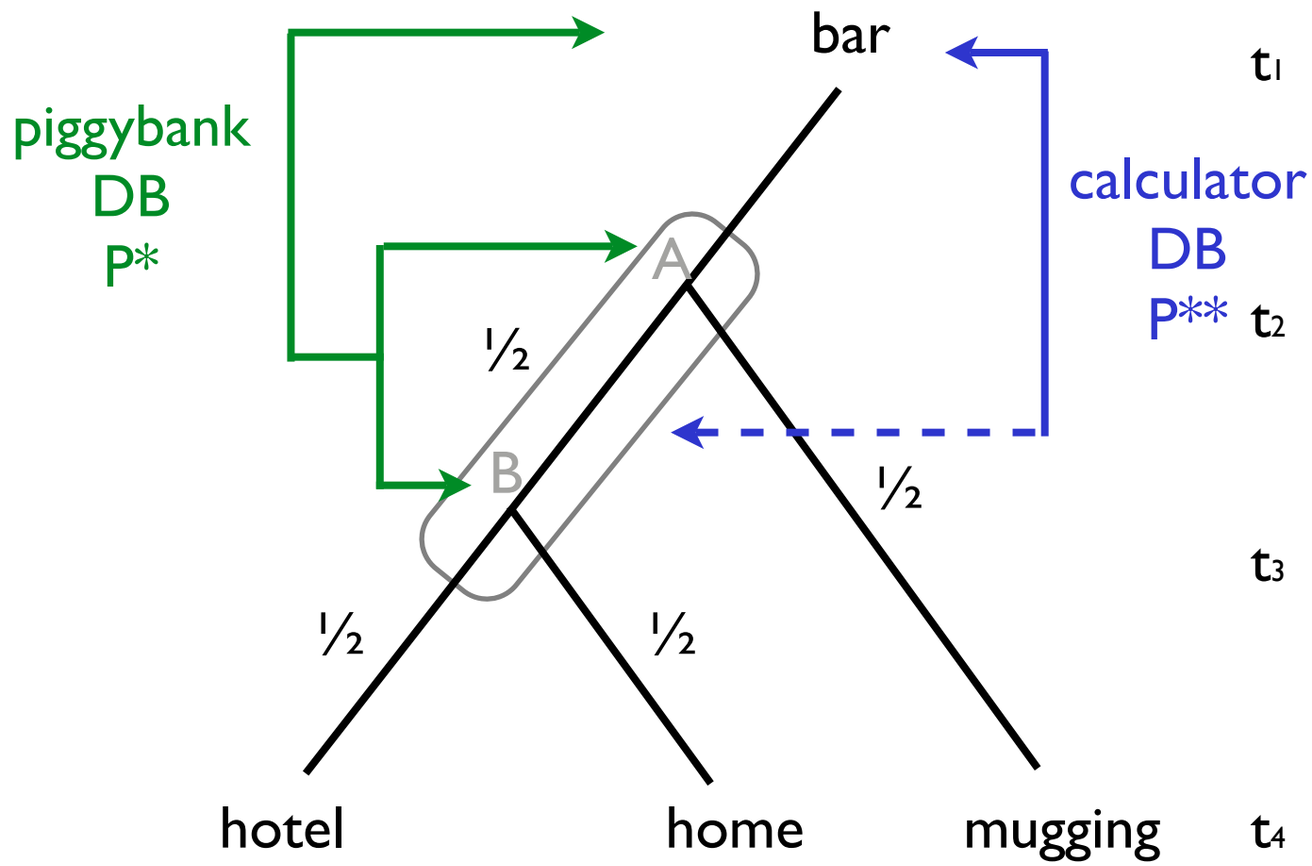
- During his period of absentmindedness, the passenger must bet such that his *total expectation* is consistent with his belief state at the bar, e.g.
- Since the probability of T is still 1/2 at A, but is 0 at B, he must bet at A and B *as if* $P^*(T) = P_1(T|A)P_{Ab}(A) + P_1(T|B)P_{Ab}(B) = 1/2(2/3) + 0(1/3) = 1/3$
- This is a bet *as if* result - should P^* represent P_2 and P_3, however? Does betting here *measure belief*?
- ***No**, because actions and beliefs have come apart.*

The Absentminded Driver

- We can see this more clearly if we remind ourselves of Ramsey's maxim that when we measure degrees of belief we are interested only in the "measurement of belief *qua* basis of action."
- Since actions may be repeated when an agent is absentminded, we must understand the relationship between action and repetition in order to discover rational degrees of belief.
- But not all actions respond to repetition the same way.
- If I put coins into a piggy bank, the number of coins accumulates monotonically.
- The assumptions which generated P* are that bets behave like coins in a piggy bank, *they accrue monotonically*.
- In fact, we might even describe P* as characterizing betting behavior if the bettor and the bookie agree at the bar to store their bets in a piggy bank. After their absentmindedness is over, they crack open the piggy bank, and add up the bets inside, making payoffs as appropriate.

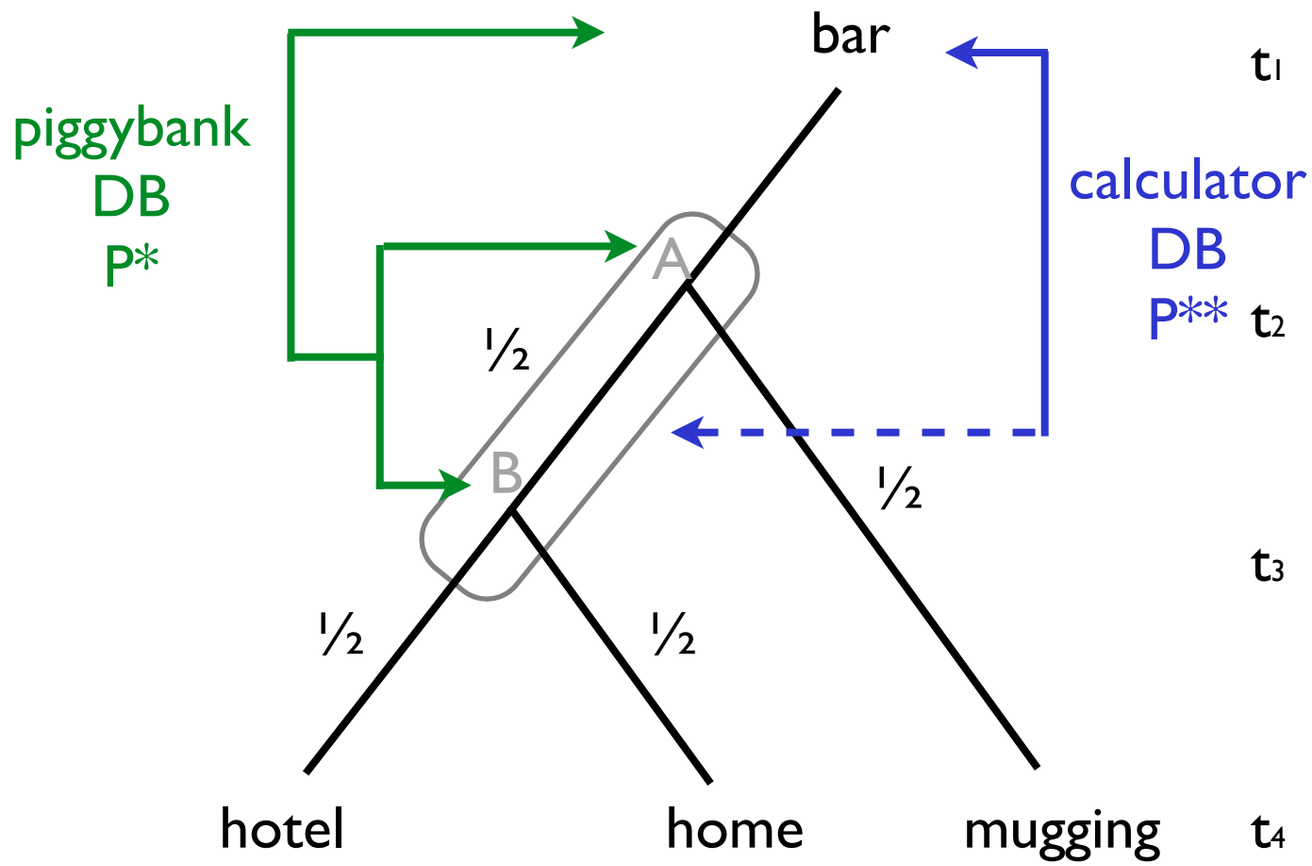
The Absentminded Driver

- Not all actions produce effects which accrue monotonically.
- Consider the example of turning of the stove. If I turn the stove off 5 times, it is no "more" *off* than if I turn the stove off once.
- Or what about the flipping of a light switch? The end effect of a number of light switch flippings is different depending upon whether it is flipped an even or an odd number of times.
- We might probe the appropriate *as if* behavior of an absentminded agent by placing bets against him using a mechanism which stores bets in a manner which reflects the changes in the outcome of an action of interest when it is repeated.
- For example, bets might be stored on an old fashioned calculator, with only one memory compartment. Only the last bet made will survive the period of absentmindedness, all others will be written over if a new bet is placed.
- Storing bets on a calculator mimics the behavior of actions like turning off the stove, which produce the same effect no matter how many times they are performed.



$P^* \neq P^{**}$

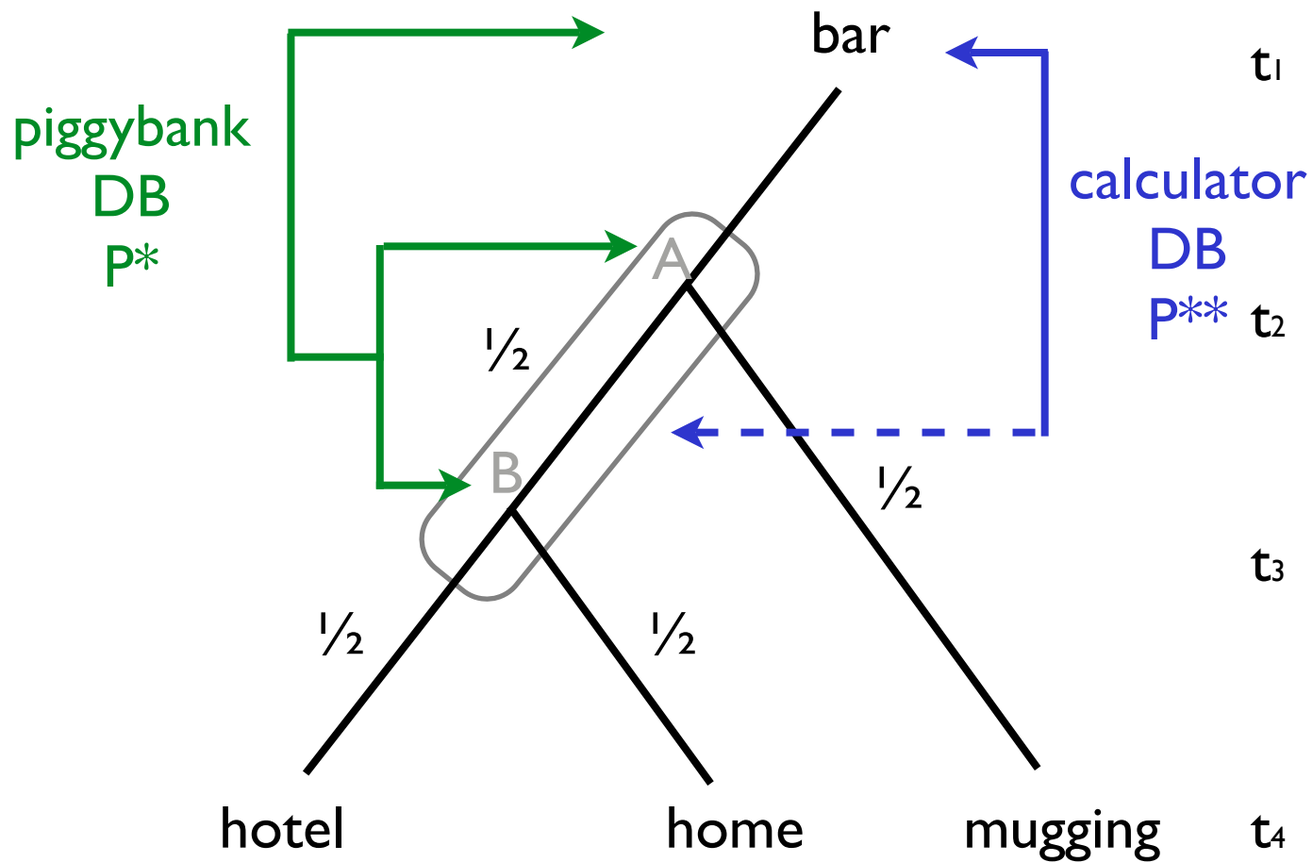
- If the absentminded passenger offers odds to a cunning "calculator" bettor, it is easy to show he must bet as if $P^{**}(T) = P_1(T) = 1/2$.
- Notice that nothing prevents the absentminded passenger from offering odds to both types of bettors simultaneously. At the bar, he agrees to store bets made by the green bettor in a piggybank and those made with the blue bettor in the old calculator.
- But since $P^* \neq P^{**}$, neither can be a simple-minded characterization of the passenger's beliefs!



$P^* \neq P^{**}$

- The trick here is to see that P* and P** represent betting behaviors *calculated* by a rational agent from

  1. his knowledge of the relative probability of all situations amongst which he is absentminded,
  2. his knowledge of the causal structure of the world, and
  3. his knowledge of the bet storing procedure (*mutatis mutandis*, the action of interest to the agent in the real world).



$P^* \neq P^{**}$

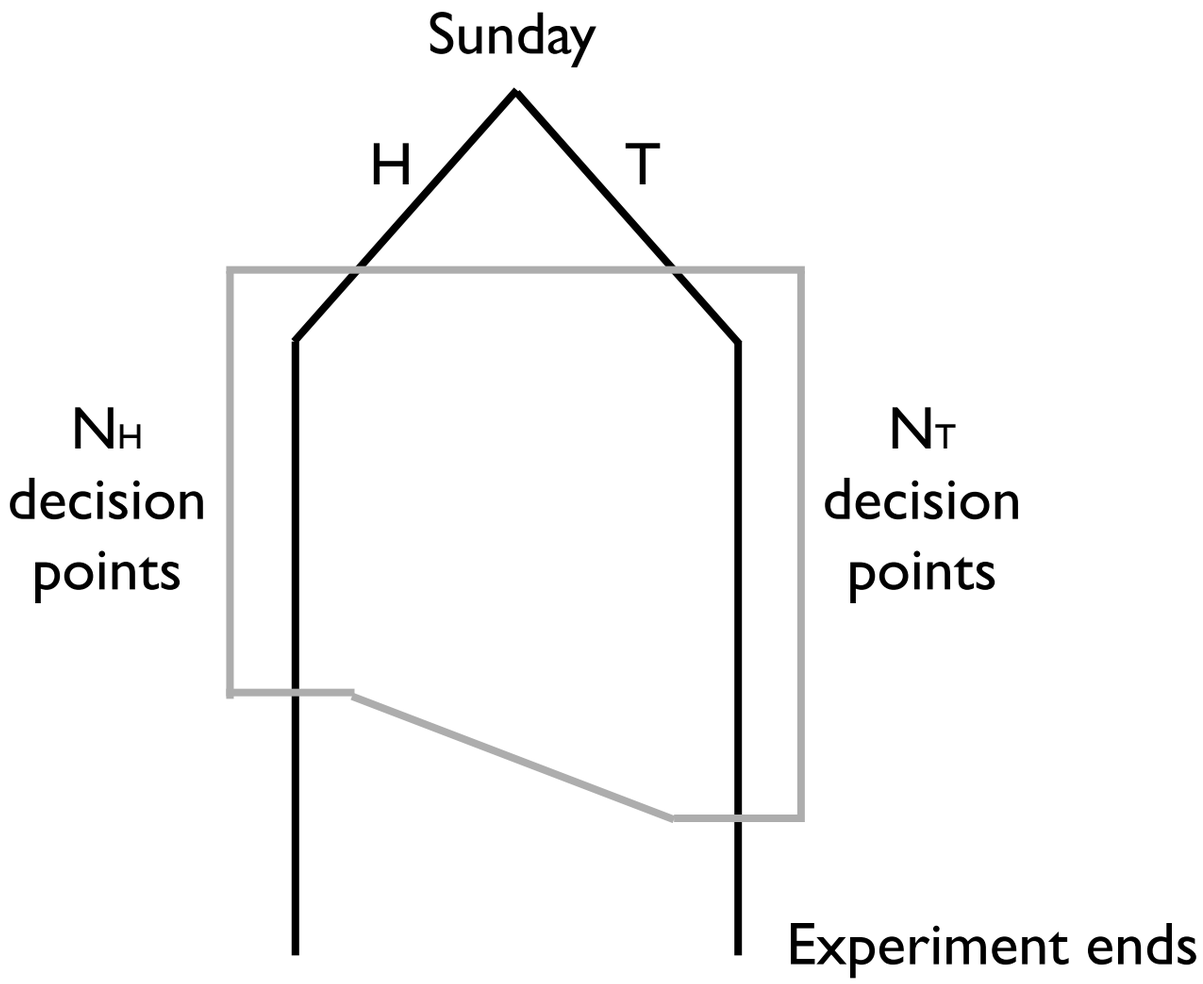
- Project: *uncover the general formula for absentminded betting*, and look for the component which looks most like a belief state. Use this as the rational recommendation for absentminded belief (and actions).
- This will tell us the demands consistency places on an absentminded agent.
- Let's look at another example in order to see how this works.

Outline of the Talk

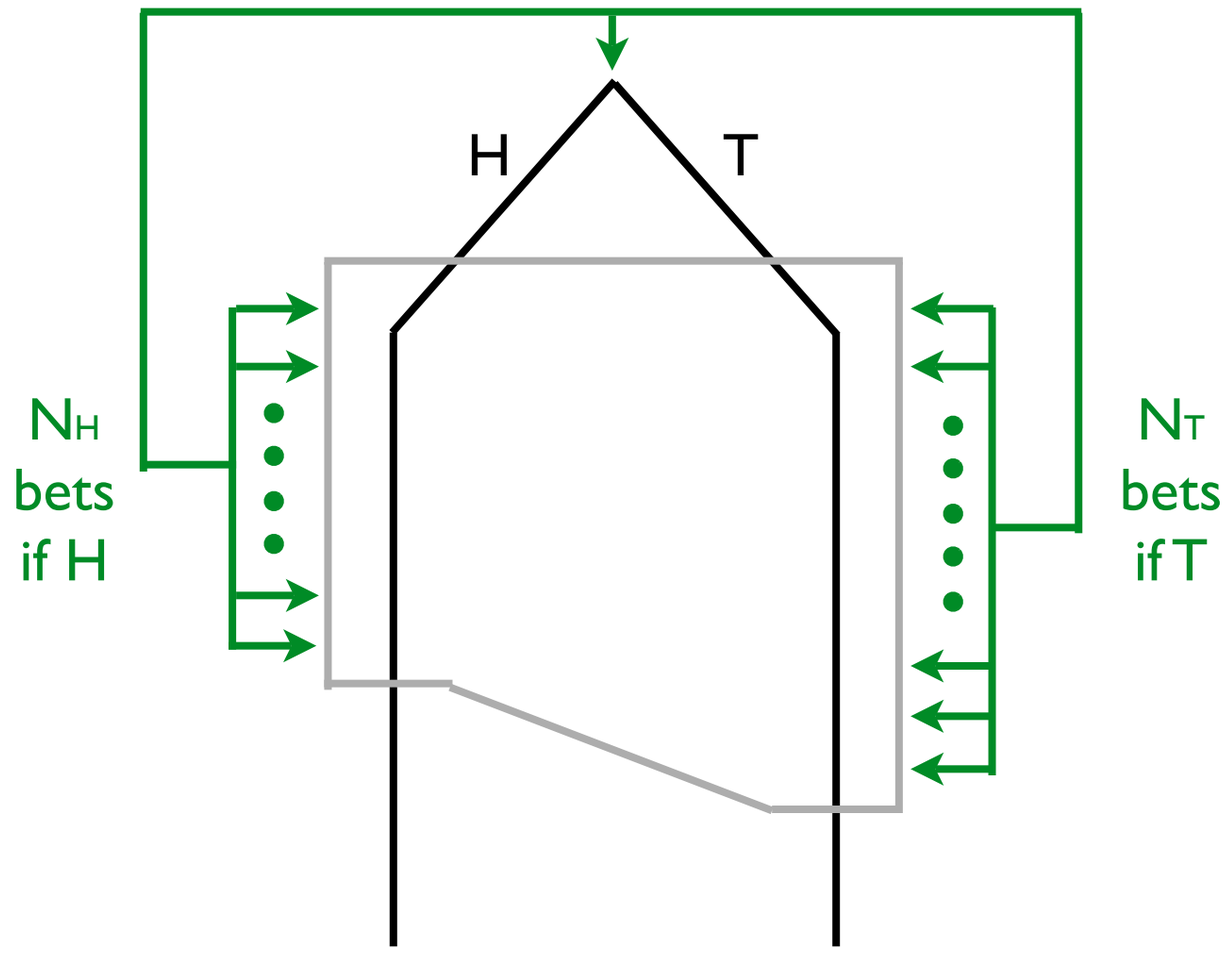
1. Introduction: *Rationality for bounded agents*
2. Dutch Book Arguments
    a. What are Dutch book arguments?
    b. Two types of objection to *diachronic* Dutch book arguments
3. Modeling forgetful agents
    a. Decision problems
    b. Imperfect recall
    c. Absentmindedness
4. The Skyrms strategy for diachronic Dutch books
5. Imperfect recall: The spaghetti dinner
6. Absentmindedness
    a. The challenge: *Action and belief come apart*
    b. The absentminded driver
    c. Sleeping beauty

Sleeping Beauty

- The absentminded driver is an example where absentmindedness encompasses a single history, which the agent may exit with various probabilities. This means that nodes in a single history are assigned different probabilities.
- Sleeping beauty (Elga, 2000) is an example of absentmindedness which includes several histories, but within each, the probability of reaching the next node is always one. This means that nodes in a single history are always assigned equal probabilities.
- Beauty is the subject of an experiment. A coin is flipped Sunday night and she is put to sleep. If the outcome is heads, she is awoken on Monday, then her memory is erased, and she is put to sleep until Wednesday. If the outcome is tails, Beauty is awoken twice, once on Monday, and a second time on Tuesday, but after each day her memory is erased before she is put to sleep. Then she is awoken Wednesday.
- More generally, a binary event H or T (= ~H) occurs to which Beauty assigns probabilities α and $1 - \alpha$, and if H occurs she is awoken N_H indistinguishable times while if T occurs she is awoken N_T indistinguishable times.

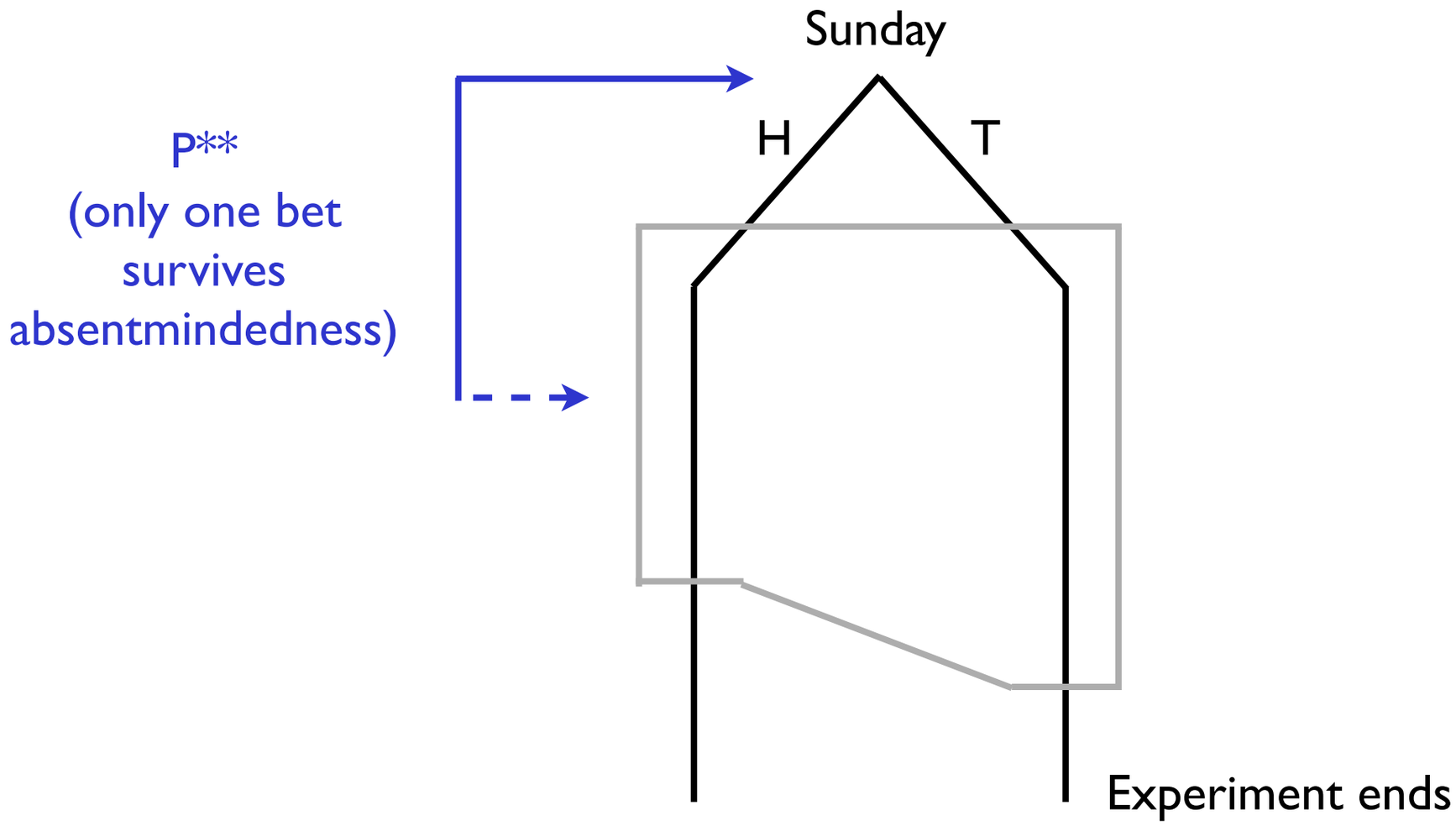


- Beauty experiences N_H indistinguishable decision nodes with probability α and N_T indistinguishable decision nodes with probability $1 - \alpha$.
- If she offers odds to a piggybank bettor, she must take into account that the bet will be placed N_H times if H and N_T times if T.

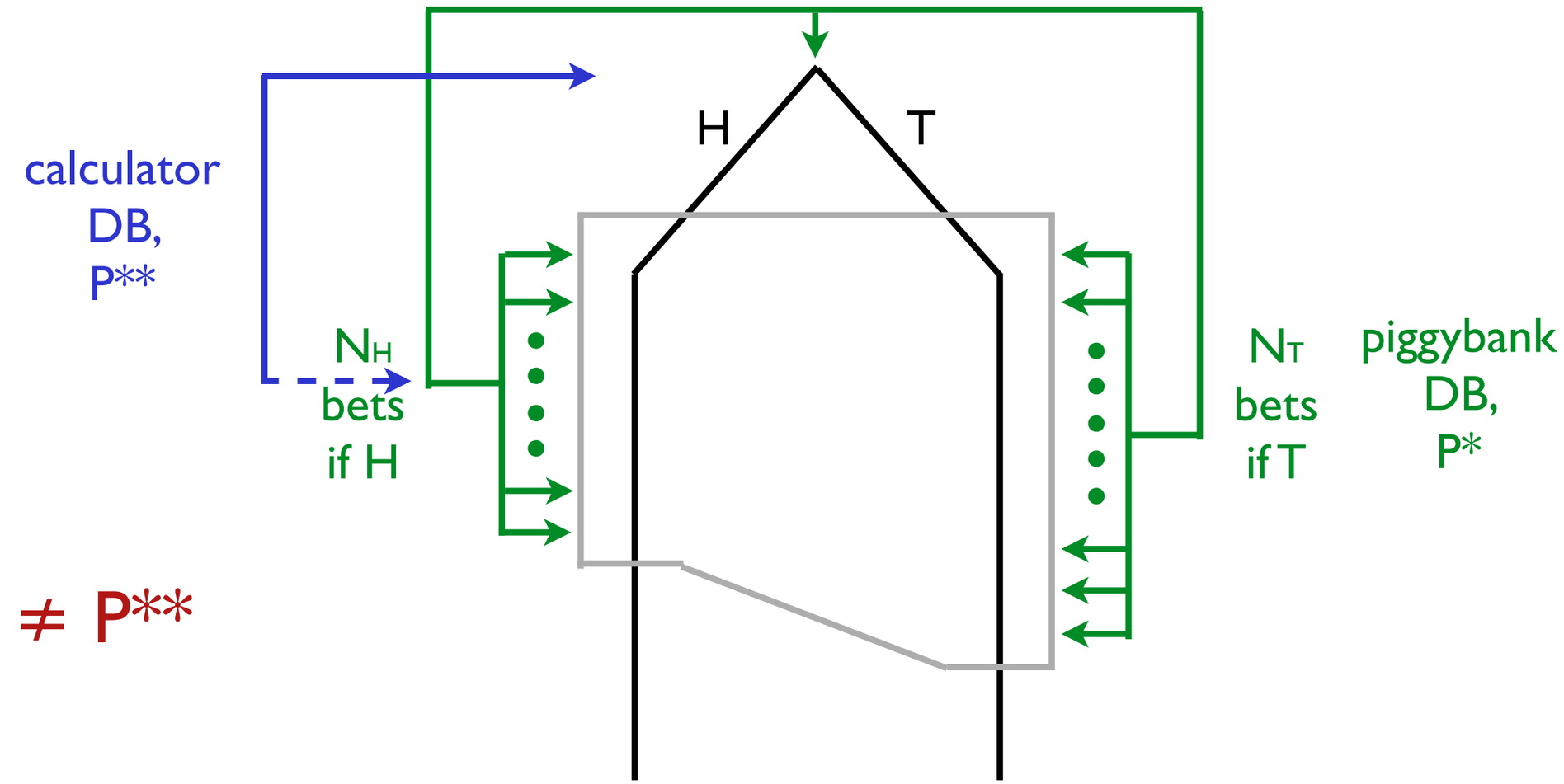


- We can show using a Skyrms-style Dutch book argument that Beauty should offer odds P* against a piggybank bettor via a straightforward expected utility calculation, e.g. for H:

$$P^*(H) = \frac{\alpha N_H}{\alpha N_H + (1-\alpha) N_T}$$



- Just as in the absentminded driver, Beauty might also offer odds to a bettor which will be stored in an old calculator. This means that only a single bet will survive Beauty's period of absentmindedness.
- It is easy to show that $P^{**}(H) = P_1(H)$



- Is there a general formula, which characterizes correct betting procedure for both Sleeping beauty and the Absentminded driver, including both piggybank and calculator bets?
- The answer includes a function from decision node and betting procedure to histories across which to assess probability.

Sleeping Beauty

- Call the set of nodes across which the agent is absentminded, Z.
- Set h(x) as the set of all histories (visualize them as a "fan") which pass through node x.
- Then take $f_b(x)$ to be the set of relevant nodes for assessing the probability of some proposition *at* x, *given betting procedure* b.
- Then, the betting odds P* an absentminded agent should offer on the history $A \in \Omega$ for betting procedure b are given by

$$P^*(A) = \sum_{z \in Z} \frac{P_I(h(z))}{\sum_{z' \in Z} P_I(h(z'))} P_I(A | \bigcup_{v \in V} h(v))$$

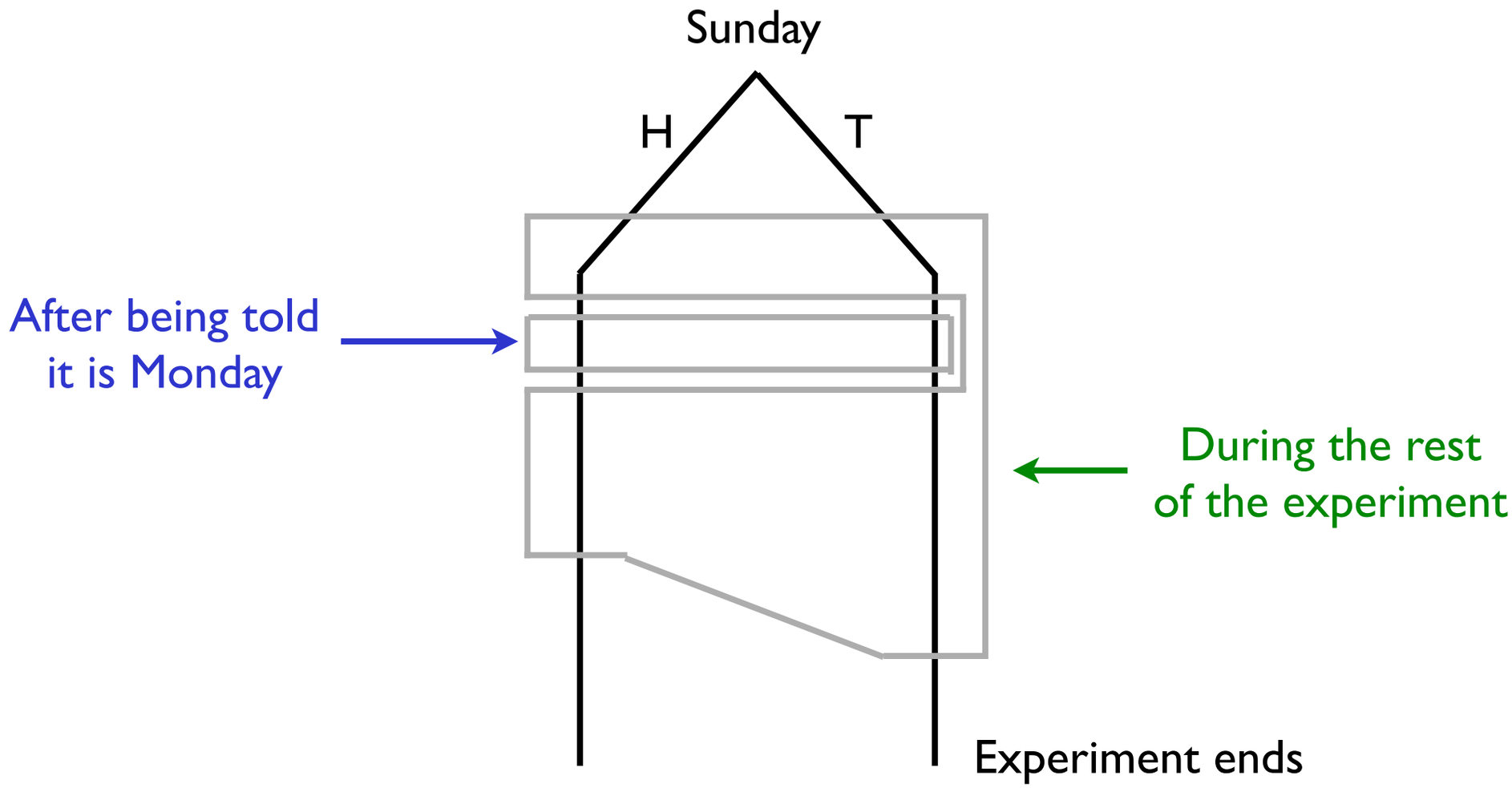
- Where $V = f_b(z)$. In the case of piggybank bets, $f_b(z) = z$. In the case of calculator bets, $f_b(z) = Z$ (the set of all absentminded nodes). This general format can also handle bets that only occur on some histories and not on others.

Sleeping Beauty: Morals

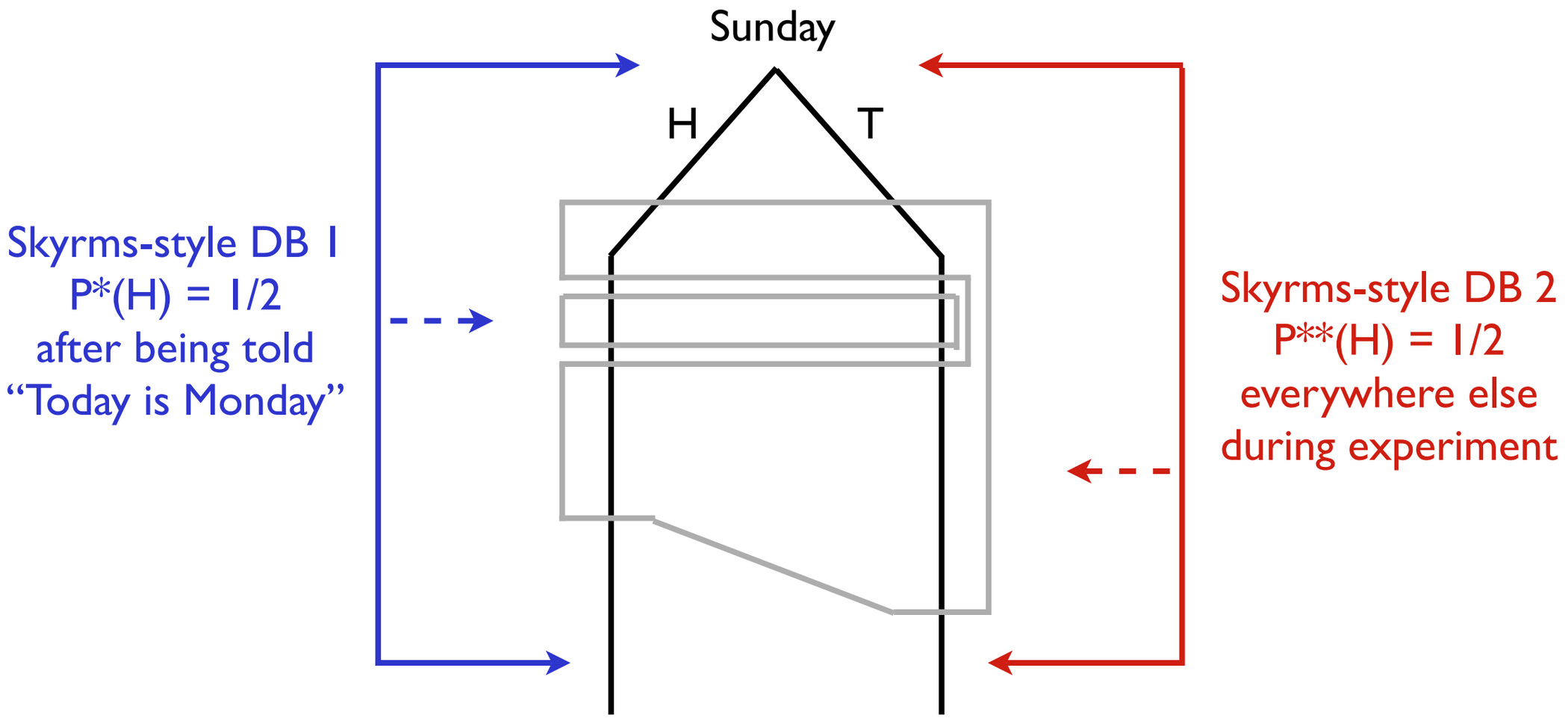
1. We can use Dutch book arguments to make the calculation of correct betting behavior for absentminded agents transparent.
2. As long as the bettor is in the same epistemic state as the bookie (i.e. has access to the same information about the world), there is nothing illegitimate or "unfair" about these bets.
3. Insofar as there are decision problems involving absentmindedness in the real world, these results apply much more readily to them than to cases of imperfect recall since the decision points of interest plausibly fall much closer together (and, consequently, are less likely to have important decision points in between them).
4. There is no guarantee that any of these results (including Lewis'!) extend to problems which are more complex or are not bracketed.

Coda: *Monday?*

- There is an important consequence of these considerations, namely that it is unclear what happens to even basic epistemic principles (e.g. *conditionalization*) in significantly more complex decision problems.
- For example, both "halfers" and "thirders" about Sleeping Beauty often agree that if she learns (i.e. is told veridically) that "Today is Monday," she should update her beliefs via conditionalization, which can be calculated with Bayes' rule.
- Call P_M her belief state after hearing "Today is Monday," then, for the halfer, $P(H) = 1/2$, but $P_M(H) = 2/3$.
- For the thirder, $P(H) = 1/3$, but $P_M(H) = 1/2$.
- Draper and Pust (2008) have tried to use this assumed application of conditionalization to provide a Dutch book in support of the thirder conclusion.

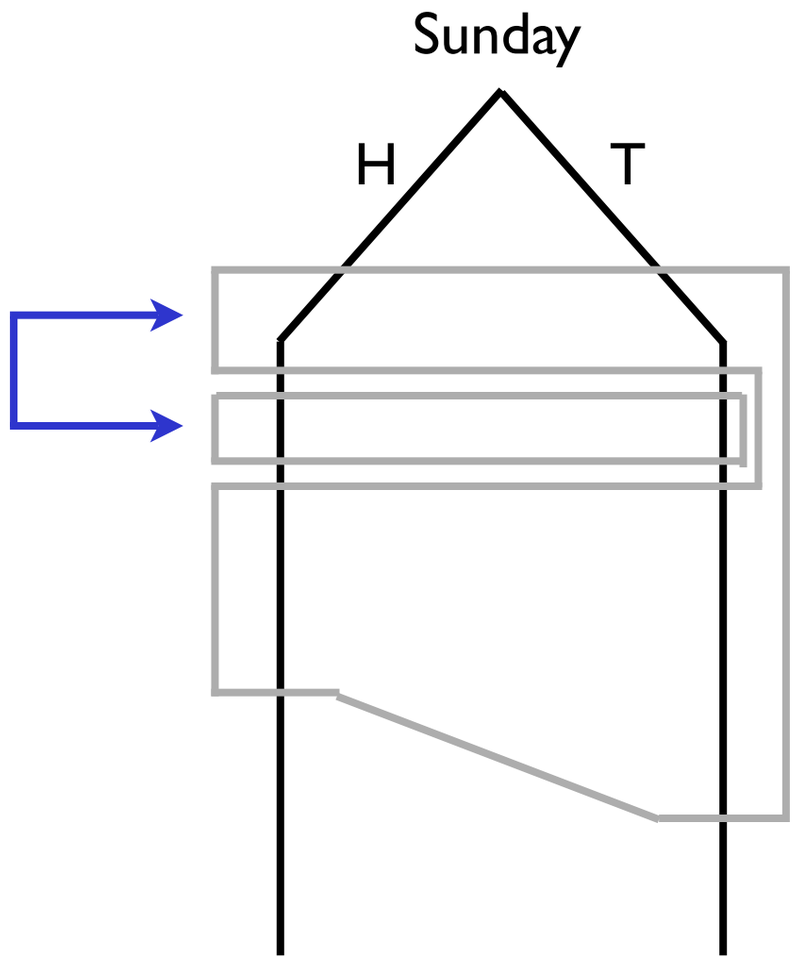


- If Sleeping Beauty is told it is Monday, she temporarily enters a state during which she is still uncertain about the outcome of the coin toss.
- However, she reenters the previous information set again after her memory is wiped at the end of the day.



- But we can use a Skyrms-style Dutch book to show that her belief in heads should be 1/2 *both* during the rest of the experiment *and* after she has been told that it is Monday!
- [Halpern (2005) makes a similar point for different reasons.]
- SO, we go wrong if we conditionalize after being told “Today is Monday” no matter where we start! How can that be?

Problem not "bracketed": agent is not epistemically secure on Monday morning, so Lewis DB does not apply



- Notice, the conditions of Lewis' Dutch book are not satisfied during the shift from Monday morning to Monday afternoon:
- In particular, the agent is not *epistemically secure* on Monday morning. If the preconditions of Lewis' DB are not met, then it does not recommend that the Beauty conditionalize when told "Today is Monday."
- Moral: *results from simple decision problems do not necessarily extend to complex decision problems!*

This research was made possible by a postdoctoral fellowship from the *McDonnell Foundation Research Consortium on Causal Learning*.

References

- Aumann, et al. (1997) "The Forgetful Passenger"
- Border & Segal (1994) "Dutch Books and Conditional Probability"
- Bradley & Leitgeb (2006) "When Betting Odds and Credences Come Apart: More Worries for Dutch Book Arguments"
- de Finetti (1937) "Foresight: Its Logical Laws, Its Subjective Sources"
- Elga (2000) "Self-Locating Belief and the Sleeping Beauty Problem"
- Maher (1992) "Diachronic Rationality"
- Piccione & Rubinstein (1997) "On the Interpretation of Decision Problems with Imperfect Recall"
- Ramsey (1926) "Truth and Probability"
- Skyrms (1987) "Dynamic Coherence and Probability Kinematics"
- Talbott (1991) "Two Principles of Bayesian Epistemology"
- Teller (1973) "Conditionalization and Observation"