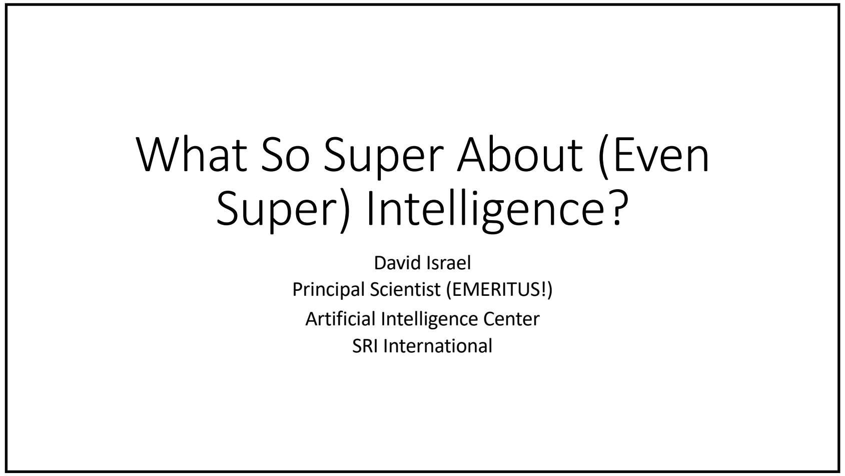
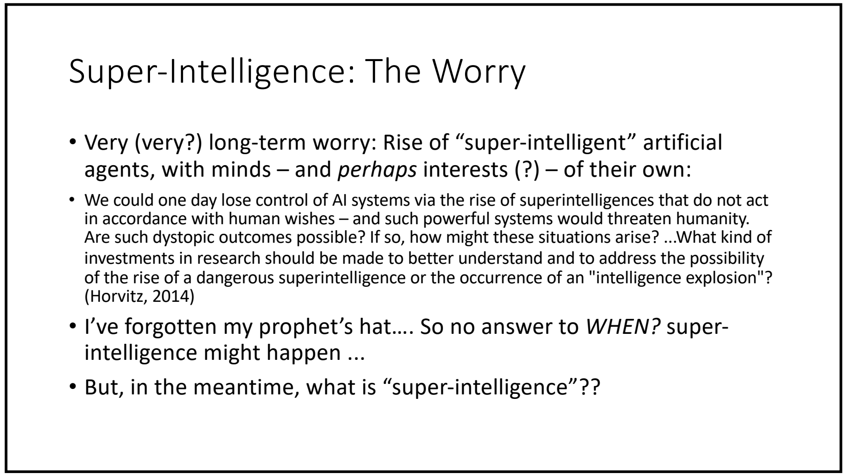


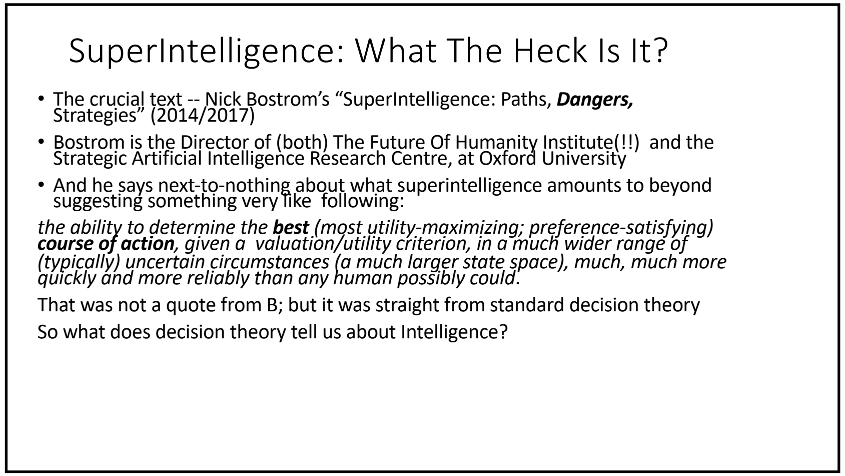
# What So Super About (Even Super) Intelligence?

David Israel
Principal Scientist (EMERITUS!)
Artificial Intelligence Center
SRI International



## Super-Intelligence: The Worry

- Very (very?) long-term worry: Rise of “super-intelligent” artificial agents, with minds – and *perhaps* interests (?) – of their own:
- We could one day lose control of AI systems via the rise of superintelligences that do not act in accordance with human wishes – and such powerful systems would threaten humanity. Are such dystopic outcomes possible? If so, how might these situations arise? ...What kind of investments in research should be made to better understand and to address the possibility of the rise of a dangerous superintelligence or the occurrence of an "intelligence explosion"? (Horvitz, 2014)
- I’ve forgotten my prophet’s hat…. So no answer to *WHEN?* super-intelligence might happen ...
- But, in the meantime, what is “super-intelligence”??



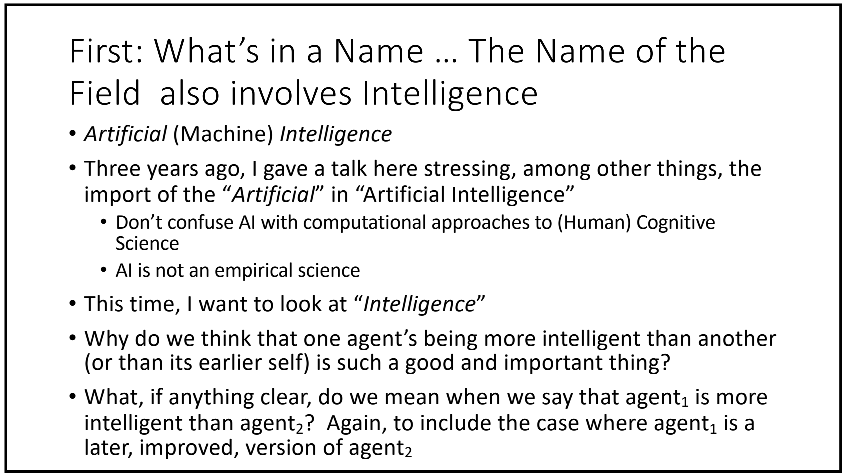
## SuperIntelligence: What The Heck Is It?

- The crucial text -- Nick Bostrom’s “SuperIntelligence: Paths, ***Dangers,*** Strategies” (2014/2017)
- Bostrom is the Director of (both) The Future Of Humanity Institute(!!) and the Strategic Artificial Intelligence Research Centre, at Oxford University
- And he says next-to-nothing about what superintelligence amounts to beyond suggesting something very like following:

*the ability to determine the **best** (most utility-maximizing; preference-satisfying) **course of action**, given a valuation/utility criterion, in a much wider range of (typically) uncertain circumstances (a much larger state space), much, much more quickly and more reliably than any human possibly could.*

That was not a quote from B; but it was straight from standard decision theory

So what does decision theory tell us about Intelligence?



## First: What’s in a Name … The Name of the Field also involves Intelligence

- *Artificial* (Machine) *Intelligence*
- Three years ago, I gave a talk here stressing, among other things, the import of the “*Artificial*” in “Artificial Intelligence”
  - Don’t confuse AI with computational approaches to (Human) Cognitive Science
  - AI is not an empirical science
- This time, I want to look at “*Intelligence*”
- Why do we think that one agent’s being more intelligent than another (or than its earlier self) is such a good and important thing?
- What, if anything clear, do we mean when we say that agent$_1$ is more intelligent than agent$_2$? Again, to include the case where agent$_1$ is a later, improved, version of agent$_2$

## What’s So Good about Being Intelligent? Never Mind Being SuperIntelligent

- What do (should) we – AI researchers – mean by *Intelligence ?*
- Not measured by a score on a standard IQ test
    - Or ability at *Jeopardy*
- Not the ability to play Chess or Go at human or even super-human levels – at least not just this or not especially this
- What we do or should mean is *Intelligence in Action*
    - *PRACTICAL INTELLIGENCE*
- And, with respect to comparisons between agents, what do or should we mean by saying: Agent$_1$ is *More Intelligent in Action* than Agent$_2$ and/or is *at least as Intelligent, but across a wider range of Activities*

## Nota Bene

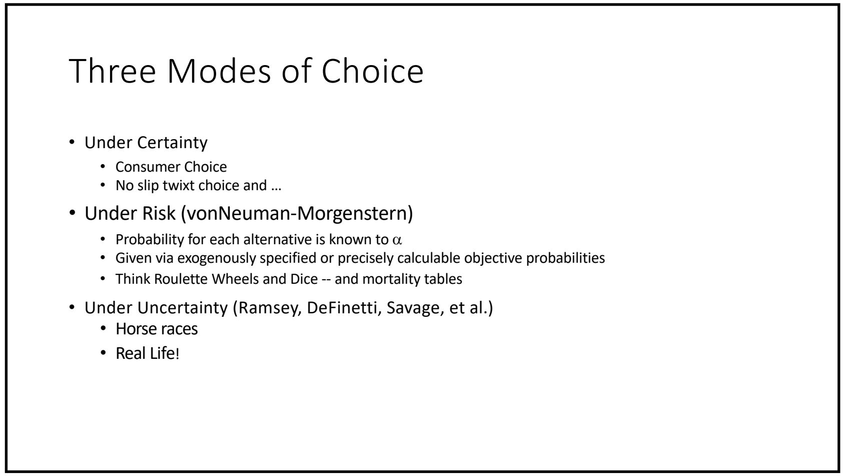
- We might well say that Deep Blue was a better chess-player than Gary Kasparov or than any human
- And that Watson was better at Jeopardy than any human
- But it would be ***crazy*** to say that Deep Blue or Watson was more intelligent than any human
- Why???

## Comparisons Of Intelligence

- Surely we must mean something that like:
- Agent$_1$ is more successful, with respect to a given set of its own purposes/goals/criteria of success — and reliably/systematically more successful -- than Agent$_2$, when they are in similar environments and faced with similar problems/challenges
- Agent$_1$ is at least as successful as Agent$_2$ – reliably/systematically – with respect to those criteria of success over a wider range of environments
- Or with respect to a wider range of problems/challenges
- Or relative to a larger set of purposes/with respect to a more encompassing set of criteria of success

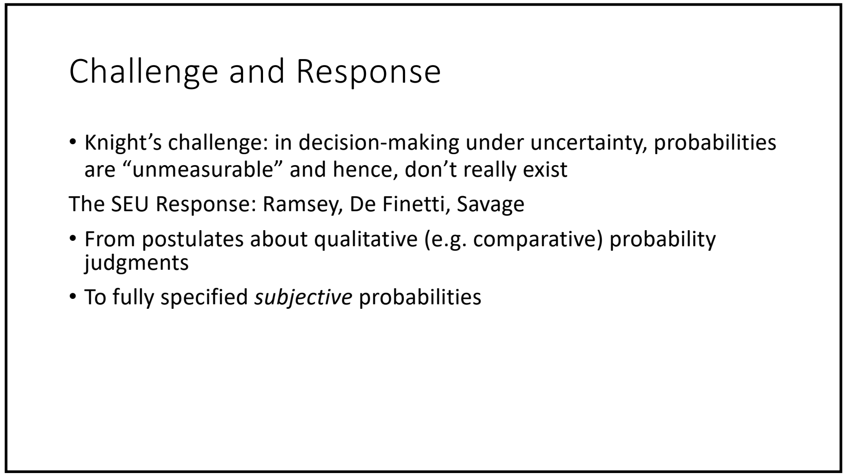
## This is Completely in Line with the (Standard) Decision-Theoretic Framework

- Agents are modeled as out to maximize their expected utility
- When in a choice/decision situation: they choose that action that, *given what they believe*, is most likely to get them more of whatever they want (or less of what they really want to avoid…)
- Or, put slightly differently: that action which *if their beliefs were true*, *would* get them more of what they want or would best promote their interests
- Really a theory of action-choice, of the *mental states that motivate and cause actions,* not directly a theory of action.
- But then what would a theory of action look like?
    - Israel, Perry, Tutiya, “*Actions and Movements*”, IJCAI-91



## Three Modes of Choice

- Under Certainty
  - Consumer Choice
  - No slip twixt choice and …
- Under Risk (vonNeuman-Morgenstern)
  - Probability for each alternative is known to $\alpha$
  - Given via exogenously specified or precisely calculable objective probabilities
  - Think Roulette Wheels and Dice -- and mortality tables
- Under Uncertainty (Ramsey, DeFinetti, Savage, et al.)
  - Horse races
  - Real Life!

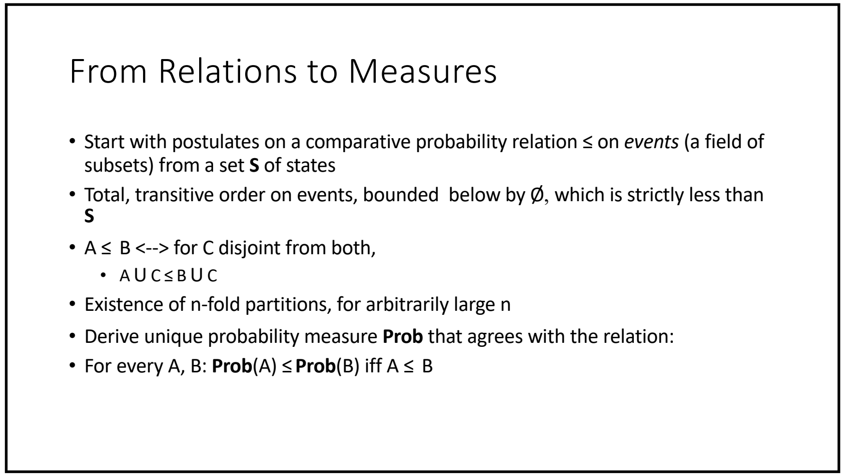


## Challenge and Response

- Knight's challenge: in decision-making under uncertainty, probabilities are "unmeasurable" and hence, don't really exist

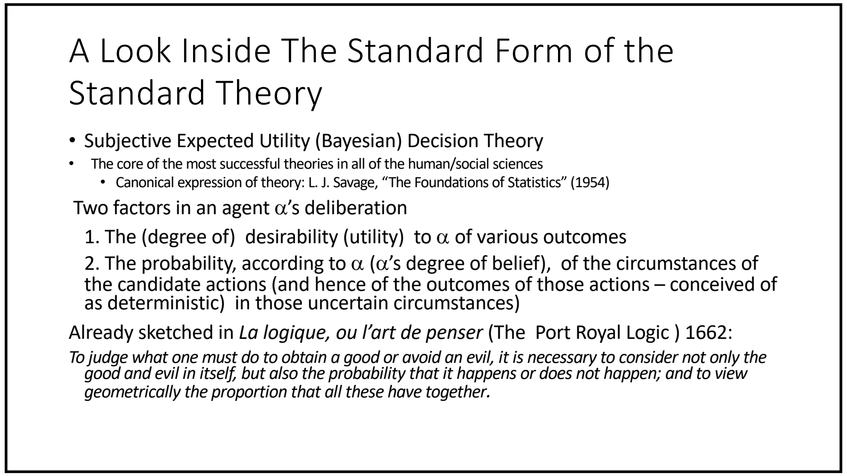
The SEU Response: Ramsey, De Finetti, Savage

- From postulates about qualitative (e.g. comparative) probability judgments
- To fully specified *subjective* probabilities



## From Relations to Measures

- Start with postulates on a comparative probability relation ≤ on *events* (a field of subsets) from a set **S** of states
- Total, transitive order on events, bounded below by Ø, which is strictly less than **S**
- A ≤ B <--> for C disjoint from both,
  - A U C ≤ B U C
- Existence of n-fold partitions, for arbitrarily large n
- Derive unique probability measure **Prob** that agrees with the relation:
- For every A, B: **Prob**(A) ≤ **Prob**(B) iff A ≤ B



## A Look Inside The Standard Form of the Standard Theory

- Subjective Expected Utility (Bayesian) Decision Theory
- The core of the most successful theories in all of the human/social sciences
  - Canonical expression of theory: L. J. Savage, "The Foundations of Statistics" (1954)

Two factors in an agent $\alpha$'s deliberation

1. The (degree of) desirability (utility) to $\alpha$ of various outcomes
2. The probability, according to $\alpha$ ($\alpha$'s degree of belief), of the circumstances of the candidate actions (and hence of the outcomes of those actions – conceived of as deterministic) in those uncertain circumstances)

Already sketched in *La logique, ou l'art de penser* (The Port Royal Logic ) 1662:

*To judge what one must do to obtain a good or avoid an evil, it is necessary to consider not only the good and evil in itself, but also the probability that it happens or does not happen; and to view geometrically the proportion that all these have together.*

## B’s and D’s

- All modern forms of Decision Theory require a quantitative or graded -- or at the very least ordered -- notion of belief
  - No room, in standard accounts, for plain (flat-out) beliefs
  - Yes, there are beliefs of Probability = 1 and hence also some (the negations of the former), of Probability = 0, but:
  - A perfectly reasonable requirement is that no empirical beliefs, no beliefs as to contingent matters, should be among them
  - So, as between **P** and **not-P,** the agent “believes” them both, *to some (perhaps equal) degree*
    - Indeed, the agent must believe them both, to some degree, and these degrees are related
      - as prob(P) and 1-prob(P)
- All modern forms of Decision Theory require a graded or ordered notion of desirability
    - There are no flat-out desires (goals?), but only degrees of desirability or ***utility***

## Savage’s Version

According To L.J. Savage “The Foundation of Statistics”

- States of the world
  - Set **S** equipped with an algebra (field) $\Sigma$ of events - subsets closed under complement and (finite/countably infinite) unions
    - In Savage: we take the full powerset of **S**
    - “a state of the world is a description of the world leaving no **relevant** aspect undescribed”
- Outcomes (Results/Consequences)
  - Just a set, with a total, transitive order -- intuitively “Anything that can happen to a person”, “States of the person” , “Possible incomes of the person”
- Acts (lotteries): $\Sigma$ – measurable finite-valued Functions from States to probability distributions over Outcomes
  - The field of the preference relation
- All these are exogenous: specified as constituting the set-up of the given decision problem
  - Probabilities and utilities are derived and hence endogeneous

## Let’s Focus on the ”S” in SEU

- Bayesian Decision Theory
- Bayesianism: Two Simple Postulates -- one synchronic, one diachronic

An agent’s belief state at a time can (should?) be modeled by a (in/finitely additive) probability function that specifies the agent’s degrees of belief

An agent should respond to new evidence by conditionalizing on that evidence. This is where Bayes’ Theorem comes inp

- Plus one more postulate about $t_0$:

Any old probability function can be stipulated as modeling the agent’s initial state – its *prior*

## That’s (Pretty) Hard-Core, Radical Bayesianism

- Various softenings/extensions are possible
- Model an agent’s belief state by a convex set of probability functions
  - Upper and lower probabilities
- Generalize conditionalizing on *evidence* (Prob = 1) to an evidence partition, whose elements are all < 1 – *Jeffrey Conditionalization*)
- Put some constraints on *priors!*
- *Remember we are modeling agents or …. designing them!*
  - Rational (“regular”) probability functions: no contingent events get 0 or 1
  - Give equal credence to all possibilities amongst which the agent’s evidence does not discriminate (Laplace/Jaynes)
  - If there is good reason to believe that the agent’s evidence (at $t_0$) supports different credences for different possibilities, then align the agent’s initial (its prior--?) probabilities with what you believe to be *its* or *your* best evidence

## Why Do/Might  Such Constraints Matter?

- Because we want our agent to be successful! Even by its own lights
- And for that, maximizing its expected utility might not be enough
- Not, for instance, if its initial beliefs, its prior probabilities, are all wrong!
    - And it doesn't have sufficient opportunity to update (conditionalize)
- That is, not if its beliefs – its credences – are way wrong – don't align well with the environment(s) in which the agent actually has to operate and there is neither time nor world enough to have experience correct those false beliefs

## Constraints on Priors: Environment- and Problem-Specific Priors!

- If we're creating agents (remember we're AI researchers), intended to operate in a certain range of environments and to execute certain tasks or engage in certain activities, then we should build-in to our agents' priors what *we* know or believe about that range of environments
- Just as we should build-in reliable perceptual mechanisms that generate largely reliable evidence on which to update the priors
- And crucially, connect the outputs of those perceptual mechanisms to action-schemes that are likely to be utility-promoting in the given range of environments
- In this way, we may (?) be partially replicating what Mother Nature does for species via evolution by natural selection.

## Decision Theory in Action: Frogs vs Flies!

- A certain pattern of irradiation on a frog's eyes indicates, in a certain range of environments, that there is a fly (or some other flying morsel) in the vicinity – flying in a certain direction relative to the frog
- That pattern also causes the frog to execute an orient-and-attack response
- It flicks out its tongue in the indicated direction
- The pattern also indicates distance, but that plays *no* role in controlling the frog's behavior. Frogs don't do depth
- "What the Frog's Eye Tells the Frog's Brain", Lettvin, Maturana, McCulloch, Pitts, 1959

## Mother Nature Must Love Frogs

- The behavior – orient and attack – is very cheap, energy-wise; so false positives or tongue misfires are of little concern
- And flies tend not to fly in fairly tight little convoys of two or three. When they do, frogs strike out, fruitlessly, in  directions midway between them
- And because Mother Nature also loves flies, she made zillions of them – so false negatives are of no account, to the frog
- All in all, with respect to predation, the frog has it pretty easy!

## Frogs vs. Lions!

- Lions hunt mostly in “packs” or family groups. Most such predation is performed by the females.
- But male lions do sometimes hunt solo
- When they do, they must be very careful in their choice of prey
- Something big enough to provide a good meal, but not too big “experienced”; healthy, but not too healthy
- Attacks are extremely expensive energy-wise
- Failed attacks, if carried on too long, can be — literally! – deadly for the lion. Lions have very little stamina. (Heart is 0.45% of body weight.)

## When it comes to Predation

- Lions face greater, more complex, challenges than do frogs
- The state space they have to deal with is – while disjoint from the frog’s – much larger
- And their (predation-specific) action repertoire is also much larger
- So lions must be more … intelligent than frogs to succeed at predation

## Back to the simple case of the frog

- Consider the brain-state of the frog that is caused by the particular pattern of ocular irradiation
- That state, like it’s peripheral immediate ancestor, is outward- and (very immediately) backward-looking : it indicates (signifies/means) that there is (was, a very very short time ago) a fly in a certain direction relative to the frog
- But it is also “forward- looking”, that is, behavior-causing and controlling
- And, if all goes well (and, of course, as intended by Mother Nature), the “vectors” of the backward-looking and “forward-looking” causal forces (one from external source to brain; the other from brain to actuation of tongue) meet at the fly!

## And the Noble Lion?

- Much more complicated sequence of ocular arrays is required as a trigger to act.
- The motivating perception only triggers action when the lion is hungry
- Much more complicated and energy-intensive behavioral response
- And the lion must, in some way, track its own energy-usage and make a complicated cost-benefit analysis as to how long to continue the attack
- That is what makes the lion noble!

## Finally, The Question: Why Do We and Why Should We Value Intelligence?

- What function does intelligence play in the successful functioning (actions/behavior) of agents?
- Its main function is to guide behavior
- In particular, it is a means to the enablement and ultimately the production of *behavioral complexity* in response to *environmental complexity*
- So, it comes in handy when environments are too complex, *in ways relevant to the success of the agent*, to be handled by simple (e.g., automatic) behavioral responses, even to reliable perceptual input

## Environmental Complexity

- Complexity is best thought of as variability or heterogeneity
- SO in the case of environmental complexity, it can be usefully measured, at least for a start, by the size of the relevant state space that the agent faces
- But it is usually (almost always!) a mistake to think of the agent as facing one (REALLY BIG) state space throughout its life
- Think rather of problem-specific variability

## Environmental Complexity (cont.)

- My state space is bigger than yours!
- Or: my problems are harder or at least bigger than yours!
- We're modeling variability in the environment by **S,** the set of relevant possibilities facing the agent with a given decision-problem
  - Let's ignore the possibility of taking different fields from a given background S (for a given agent) and just go with the full powerset
- So, in comparing the environmental complexity of two different agents, all we have to go on are
  - Cardinality (Size) of S
  - And where available, subset relations between subsets of a given set
  - If $\mathbf{S_1}$ is a subset of $\mathbf{S_2}$, then $\mathbf{S_2}$ is more complex than $\mathbf{S_1}$

## Sheer Size Seems Awfully Simple-Minded

- The case to consider is where we have two disjoint but (roughly) equi-cardinal state spaces, for two distinct agents facing distinct decision-problems
- But with – let's stipulate – roughly equal-sized (problem-specific or even general) behavioral repertoires (actions available to them in the given decision problems -- or independently of problem, as part of their *architecture*) and
- Roughly equal size problem-specific outcome sets
- Something to think about: can we compare the state spaces available to an agent of a given *complex* type (say, *homo sapiens*), in a specific small-world decision-problem with those available to an agent of a different, less complex type (say, *canis familiaris*)?

## Behavioral Complexity

- How many different actions, modes of behavior, does the agent have in its repertoire?
- Should we also think of this in a problem-specific way??
- This is a long story
- Consider Turing Machines: they can do five things
  - Read the symbol on the cell of a tape
  - Write a symbol on a cell
  - Move left one cell
  - Move right one cell
  - HALT!
- But they can compute any of the infinitely many computable functions, say from N → N, or from $N^2$ →N, etc,.
- Of course, it makes little sense to apply the decision-theoretic framework to TMs
- Kind of like the ant in Simon's parable in *The Sciences of the Artificial*

## First-Order Complexity

- A TM can track the state of its (rather special) environment perfectly
  - What symbol, if any, is on the cell of the tape at which it is now looking
- It will then do exactly what it has been *programmed* to do, given that fact about its environment *and its internal state.*
- SO it is capable of tracking and reacting (appropriately, given its program) to a range of rules of the form: If the environment is in state S, then if you're in internal state P, do ..
  - Where following the rule often involves a change of internal state
- And this can be a large set of rules (its program can be arbitrarily, but finitely long)

## Second-Order Complexity

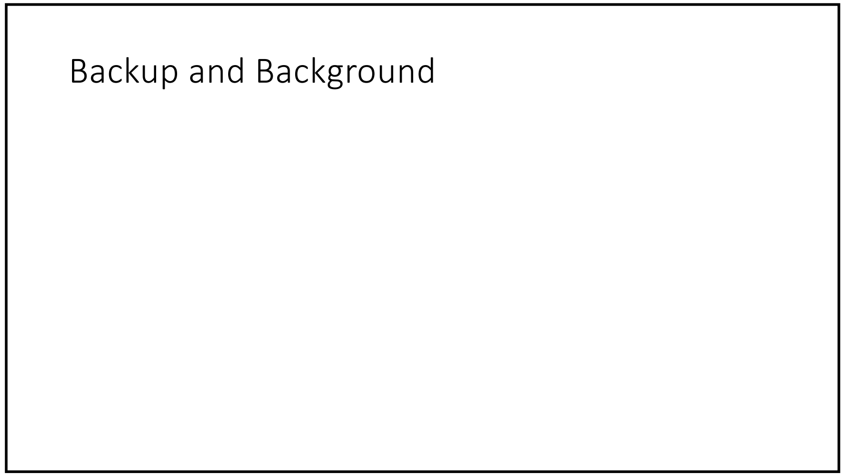
- But a (standard) TM cannot change its program – its perceptual-behavioral profile. It cannot learn
  - Of course, given its special environments and its single purpose, it doesn't have to
- An agent that can learn can change its behavioral response to given environmental conditions
- It can learn that, relative to its goals/utilities, it would be better if it executes $\text{action}_2$ rather than $\text{action}_1$ when it believes (its current credence has it that) the environment is in such-and-such a state, a state that, given the same goals/utilities, used to motivate $\text{action}_1$
- This is *not* a matter of developing new behaviors in the agent's basic repertoire
- And it is not necessary that a learning agent start off with a larger repertoire of basic behaviors than a non-learning agent

## Back to SuperIntelligence!

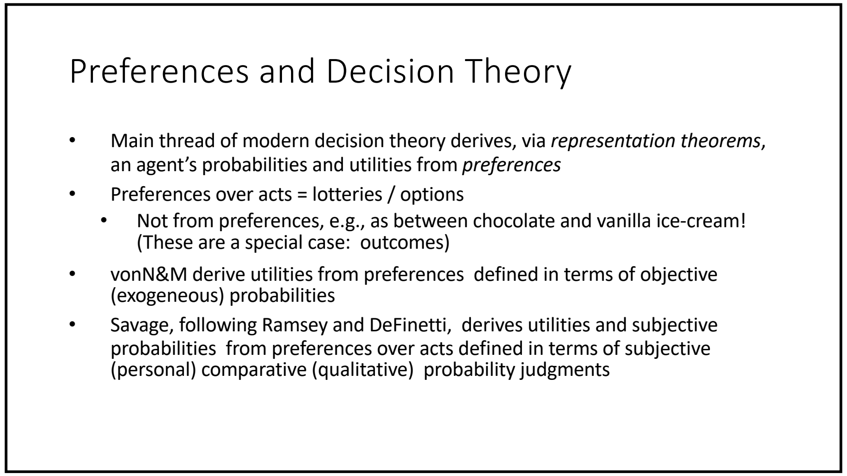
- I said that I would NOT talk about the threat posed by our creating systems much more intelligent than ourselves, but….
- Go back to my simple-minded statement of what SuperIntelligence would entail that these AI systems would have

*the ability to determine the **best** (most utility-maximizing; preference-satisfying) **course of action**, relative to a valuation/utility criterion, in a much wider range of (typically) uncertain circumstances (a much larger state space), much, much more quickly and more reliably than any human possibly could*.

So how worried should we be?

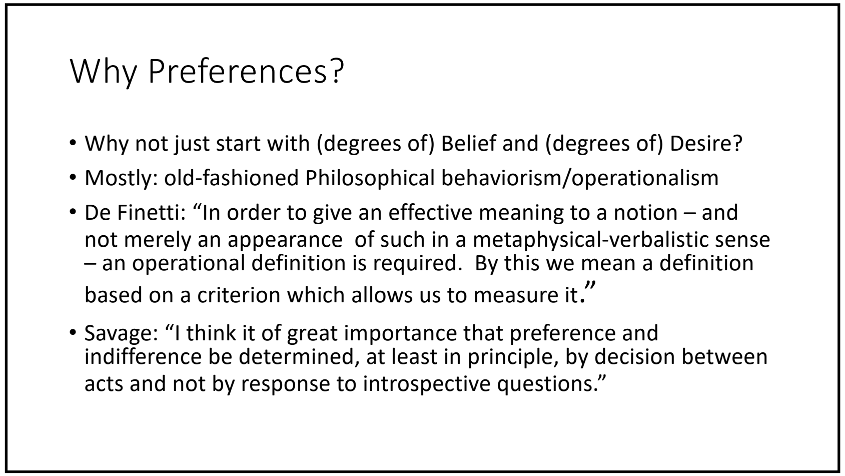


## Backup and Background



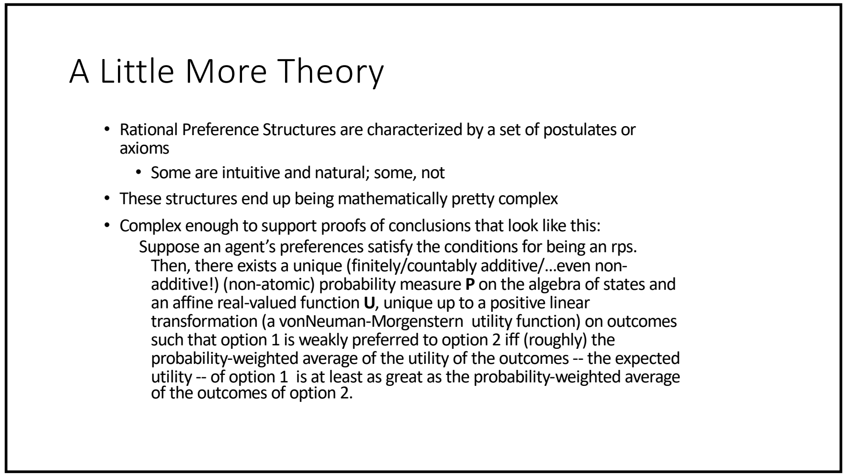
## Preferences and Decision Theory

- Main thread of modern decision theory derives, via *representation theorems*, an agent's probabilities and utilities from *preferences*
- Preferences over acts = lotteries / options
  - Not from preferences, e.g., as between chocolate and vanilla ice-cream! (These are a special case: outcomes)
- vonN&M derive utilities from preferences defined in terms of objective (exogeneous) probabilities
- Savage, following Ramsey and DeFinetti, derives utilities and subjective probabilities from preferences over acts defined in terms of subjective (personal) comparative (qualitative) probability judgments



## Why Preferences?

- Why not just start with (degrees of) Belief and (degrees of) Desire?
- Mostly: old-fashioned Philosophical behaviorism/operationalism
- De Finetti: "In order to give an effective meaning to a notion – and not merely an appearance of such in a metaphysical-verbalistic sense – an operational definition is required. By this we mean a definition based on a criterion which allows us to measure it."
- Savage: "I think it of great importance that preference and indifference be determined, at least in principle, by decision between acts and not by response to introspective questions."



## A Little More Theory

- Rational Preference Structures are characterized by a set of postulates or axioms
  - Some are intuitive and natural; some, not
- These structures end up being mathematically pretty complex
- Complex enough to support proofs of conclusions that look like this:

  Suppose an agent's preferences satisfy the conditions for being an rps. Then, there exists a unique (finitely/countably additive/…even non-additive!) (non-atomic) probability measure **P** on the algebra of states and an affine real-valued function **U**, unique up to a positive linear transformation (a vonNeuman-Morgenstern utility function) on outcomes such that option 1 is weakly preferred to option 2 iff (roughly) the probability-weighted average of the utility of the outcomes -- the expected utility -- of option 1 is at least as great as the probability-weighted average of the outcomes of option 2.

## Priors and Small-Worlds: Savage's One (!) Example

- Consider an example. Your wife has just broken five good eggs into a bowl when you come in and volunteer to finish making the omelet. A sixth egg, which for some reason, must either be used for the omelet or wasted altogether, lies unbroken beside the bowl. You must decide what to do with this unbroken egg. Perhaps it is not too great an oversimplification to say that you must decide among three acts only, namely to break it into the bowl containing the other five, to break it into a saucer for inspection, or to throw it away without inspection. Depending on the state of the egg [is it good or is it rotten?], each of these three acts will have some consequence of concern to you:
  - Break into bowl/good: 6-egg omelet
  - Break into bowl/rotten: no omelet and a waste of 5 eggs
  - Break into saucer/good: 6-egg omelet; extra dish to wash
  - Break into saucer/rotten: 5-egg omelet; extra dish to wash
  - Throw away/good: 5-egg omelet; waste of a good egg
  - Throw away/rotten: 5-egg omelet

## One Thing Leads to Another …

- Suppose you choose badly. What to do next? What to substitute? How to mollify Mrs. S?
- Rather than the single act "break into bowl", there should be several: 'break into bowl and in case of disaster, have toast', break into bowl and in case of disaster, take family to a neighboring restaurant for breakfast'.
- And, of course, you must also complexify/multiply the states and the consequences…..
- "What in the ordinary way of thinking might be regarded as a chain of decisions, one leading to the other, is in the formal description proposed here regarded as a single decision."
- But how do you know when to stop?

## Look LONG Before You Leap!

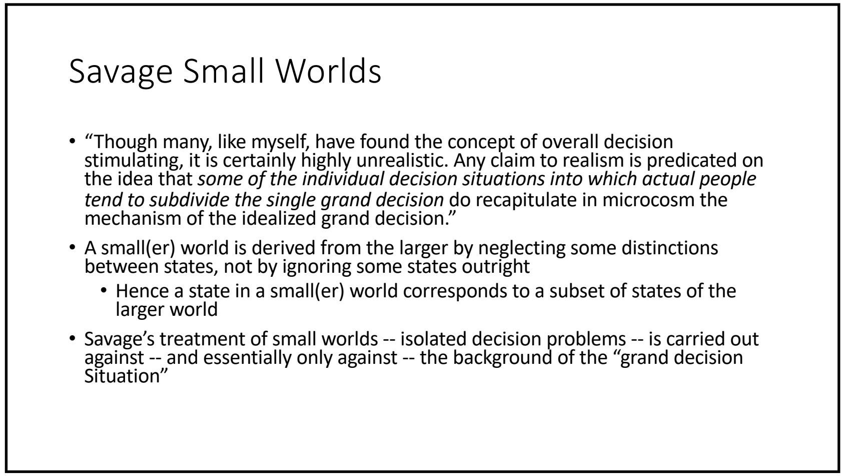
- "Carried to its logical extreme, the `Look before you leap' principle demands that one envisage every conceivable policy for the government of his whole life (at least from now on) in its most minute details, in the light of the vast number of unknown states of the world, and decide here and now on one policy".

"Making an extreme idealization, a person has only one decision to make in his whole life. He must, namely, decide how to live, and this he might in principle do once and for all."

- Life as one humungous (infinite horizon) POMDP!
  - For Savage, of order 0
- And you get no chance to explore!

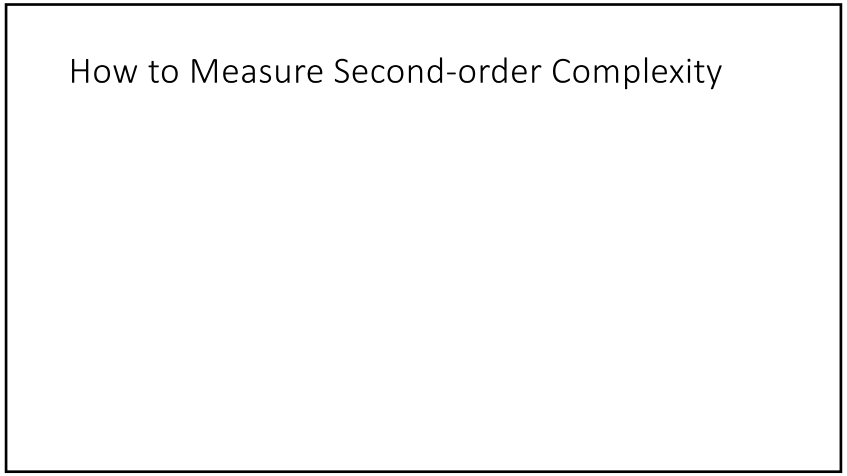
## The one big decision had better decompose into small(er) decision problems

- "Though the 'look before you leap' principle is preposterous if carried to extremes, I would none the less argue that it is the proper subject of our further discussion, because to cross one's bridges when one comes to them means to attack relatively simple problems of decision by *artificially (!!)* confining attention to so small a world that the 'Look before you leap' principle can be applied there."
- "I am unable to formulate criteria for selecting these small worlds and indeed believe that their selection may be a matter of judgement and experience about which it is impossible to enunciate sharply defined general principles…"



## Savage Small Worlds

- "Though many, like myself, have found the concept of overall decision stimulating, it is certainly highly unrealistic. Any claim to realism is predicated on the idea that *some of the individual decision situations into which actual people tend to subdivide the single grand decision* do recapitulate in microcosm the mechanism of the idealized grand decision."
- A small(er) world is derived from the larger by neglecting some distinctions between states, not by ignoring some states outright
  - Hence a state in a small(er) world corresponds to a subset of states of the larger world
- Savage's treatment of small worlds -- isolated decision problems -- is carried out against -- and essentially only against -- the background of the "grand decision Situation"



## How to Measure Second-order Complexity