Learning the Value of Value Learning

Alex John London
Department of Philosophy
Center for Ethics and Policy
Carnegie Mellon University

Aydin Mohseni
Department of Philosophy
Institute for Complex Social Dynamics
Carnegie Mellon University


What are you Bayesians supposed to do when you're uncertain about your options and values?
Good question!
Alex
Ethicist
Aydin
Epistemologist



Alex
Ethicist

Philosophers' Imprint

VOLUME 0, NO. 0
JANUARY 2017

KANTIAN DECISION MAKING UNDER UNCERTAINTY: DIGNITY, PRICE, AND CONSISTENCY

Adam Bjorndahl, Alex John London, Kevin J.S. Zollman

Carnegie Mellon University

© 2017, Bjorndahl, London, and Zollman
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License
<www.philosophersimprint.org/000000/>

Abstract

The idea that there is a fundamental difference in value between persons and things and that respecting this difference is an important moral requirement has strong intuitive appeal. Kantian ethics is unique in placing this requirement at the center of a moral system and for explicating the conditions for complying with it. Unlike challenges to Kantian ethics that focus on tragic cases that pit respect for one person against respect for another, this paper focuses on the question of how we can respect the value distinction between persons and things under conditions of uncertainty. After exploring why decision making under uncertainty is a neglected topic among Kantians and demonstrating how uncertainty challenges our ability to comply with this norm, we propose a notion of *morally insignificant risk* within a framework that allows agents to navigate real world decisions involving material benefit and some risk to dignity without violating the Kantian's core commitments. We conclude by exploring some of the challenges facing this approach.

1. Introduction

There is something compelling in the idea that there is a fundamental moral difference between persons and things, and that a significant portion of ethics is concerned with ensuring that our treatment of ourselves and others reflects this distinction. Kantian ethics places this intuitively appealing idea at the center of its moral system. A common challenge to Kantian frameworks concerns how they deal with situations in which all the choices open to an agent fail to respect the value of some individual.[1] There is, however, a different fundamental prob-

1. For instance, Marcus says, "But Kantian ethics is notably deficient in coping with dilemmas. Kant seems to claim that they don't really arise, and we are provided with no moral grounds for their resolution" (1980, pp. 125–126). In cases where each of the actions open to an agent will imperil the lives of different groups of people, Kantians face the challenge of how to decide which action to take without aggregating lives in the way a consequentialist would. Trolley problems present a similar challenge; the literature on doing/allowing, intending/foreseeing, killing/letting die, etc., represent strategies Kantians have

In *The Foundations of Statistics*, Savage articulated a foundational challenge for Bayesian decision theory.

“Making an extreme idealization, which has in principle guided the whole argument of this book thus far, 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.”

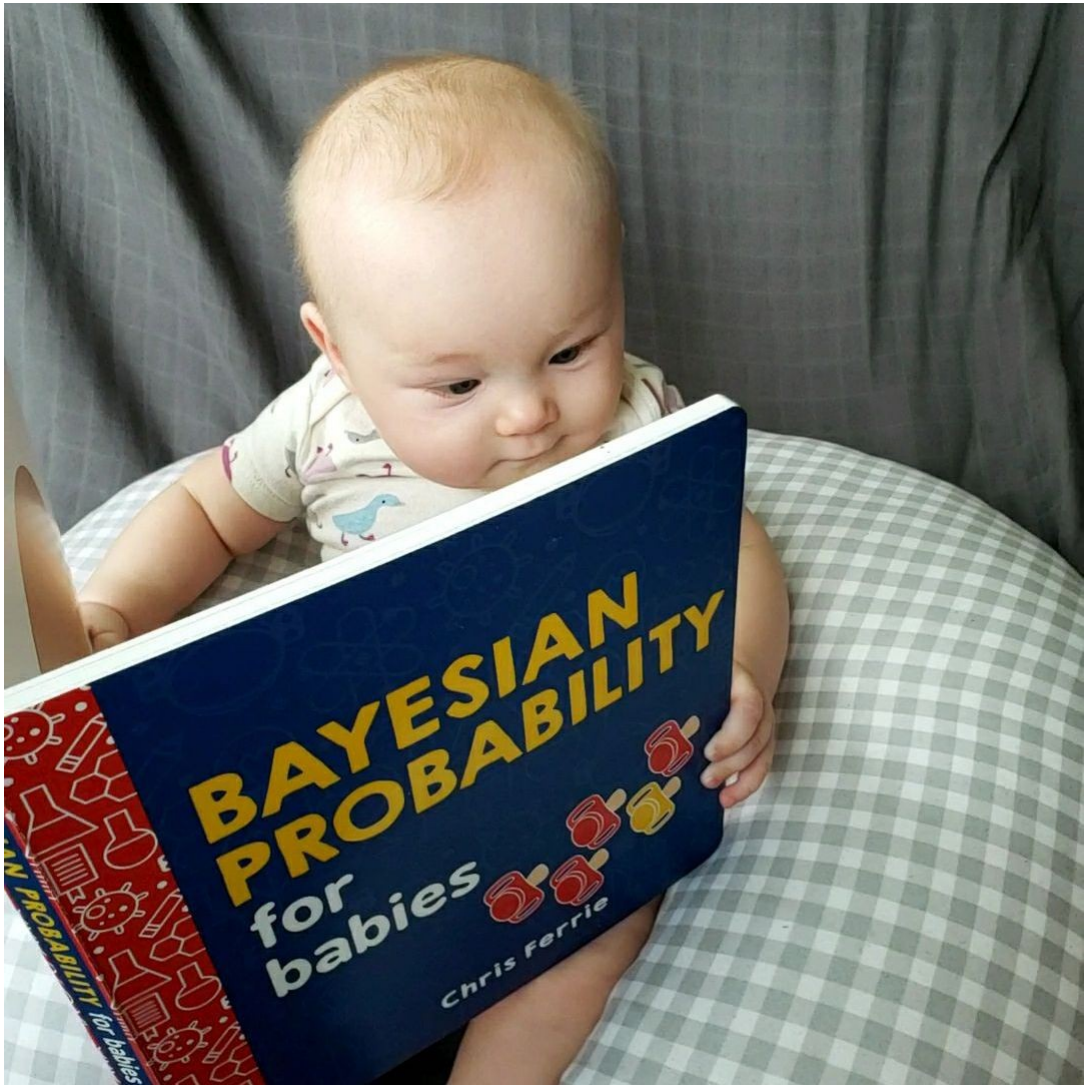
“...[this] is certainly highly unrealistic and in many contexts unwieldy.”

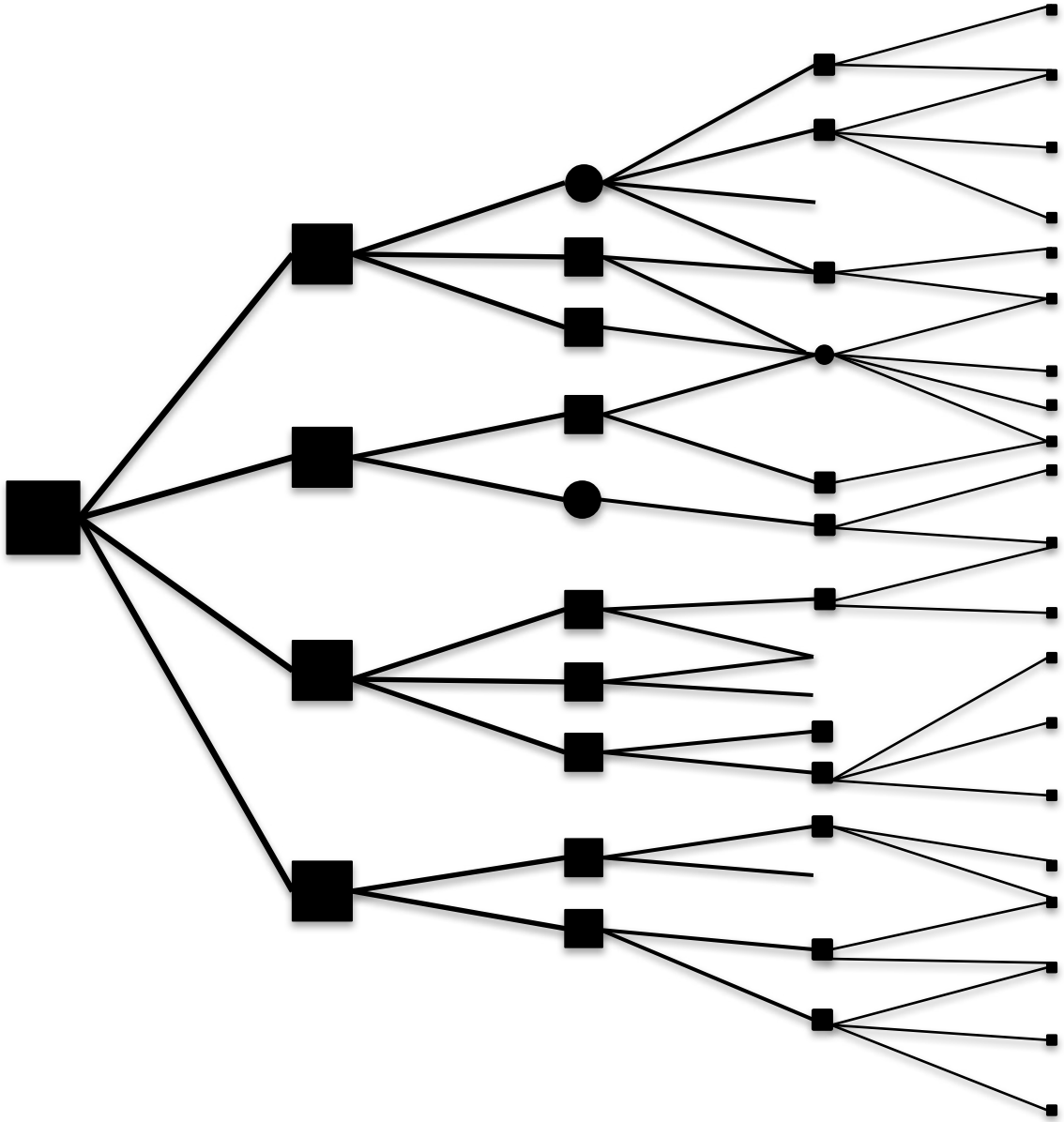
“Any claim to realism made by this book—or indeed by almost any theory of personal decision of which I know—is predicated on the idea that some of the individual decision situations into which actual people tend to subdivide the single grand decision do recapitulate in microcosm the mechanism of the idealized grand decision.”

—Savage (1951) *Foundations of Statistics*, p. 83

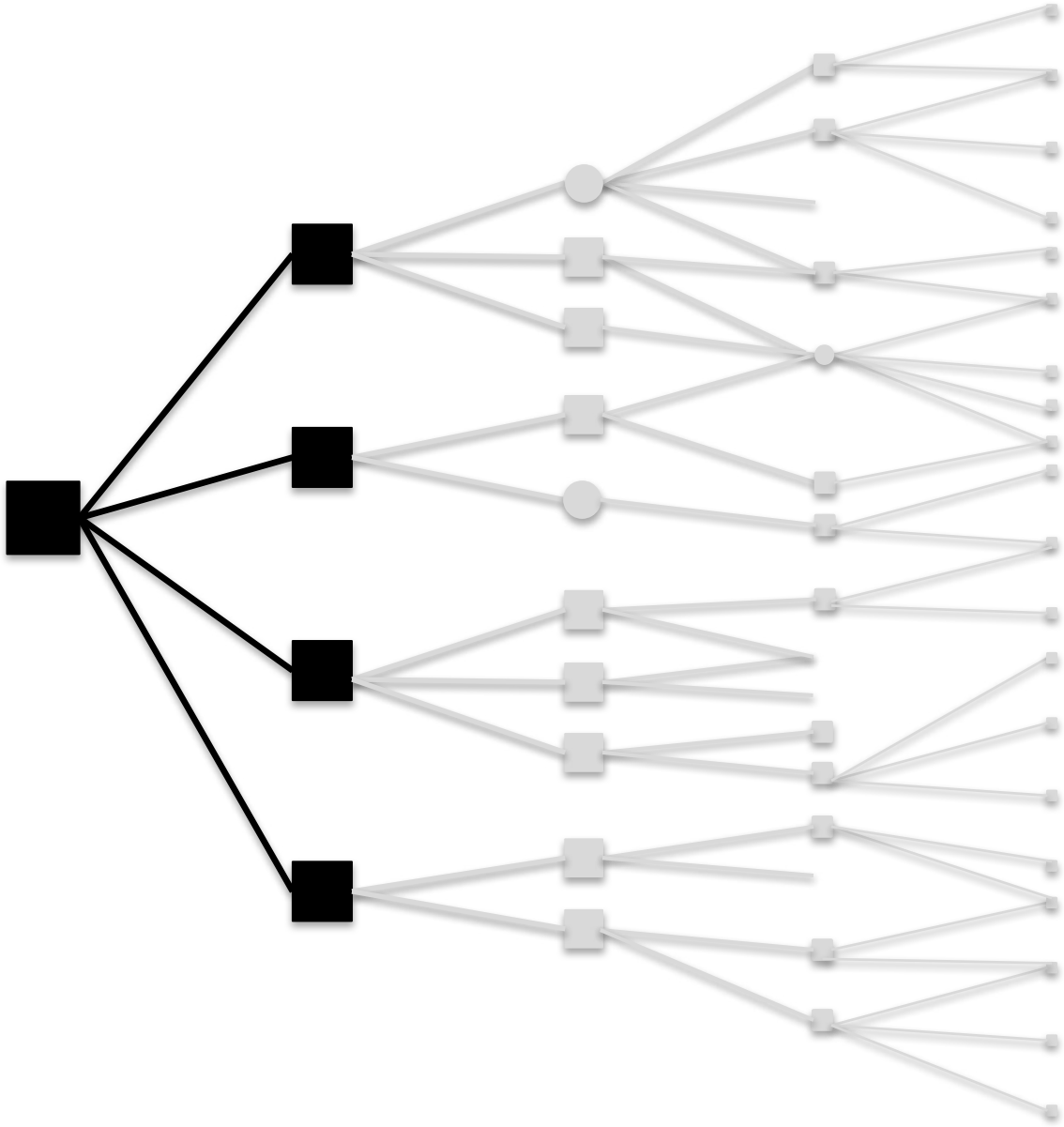
“[But] I am unable to formulate criteria for selecting these small worlds…”

—Ibid, p. 16


BAYESIAN PROBABILITY
for babies
Chris Ferrie



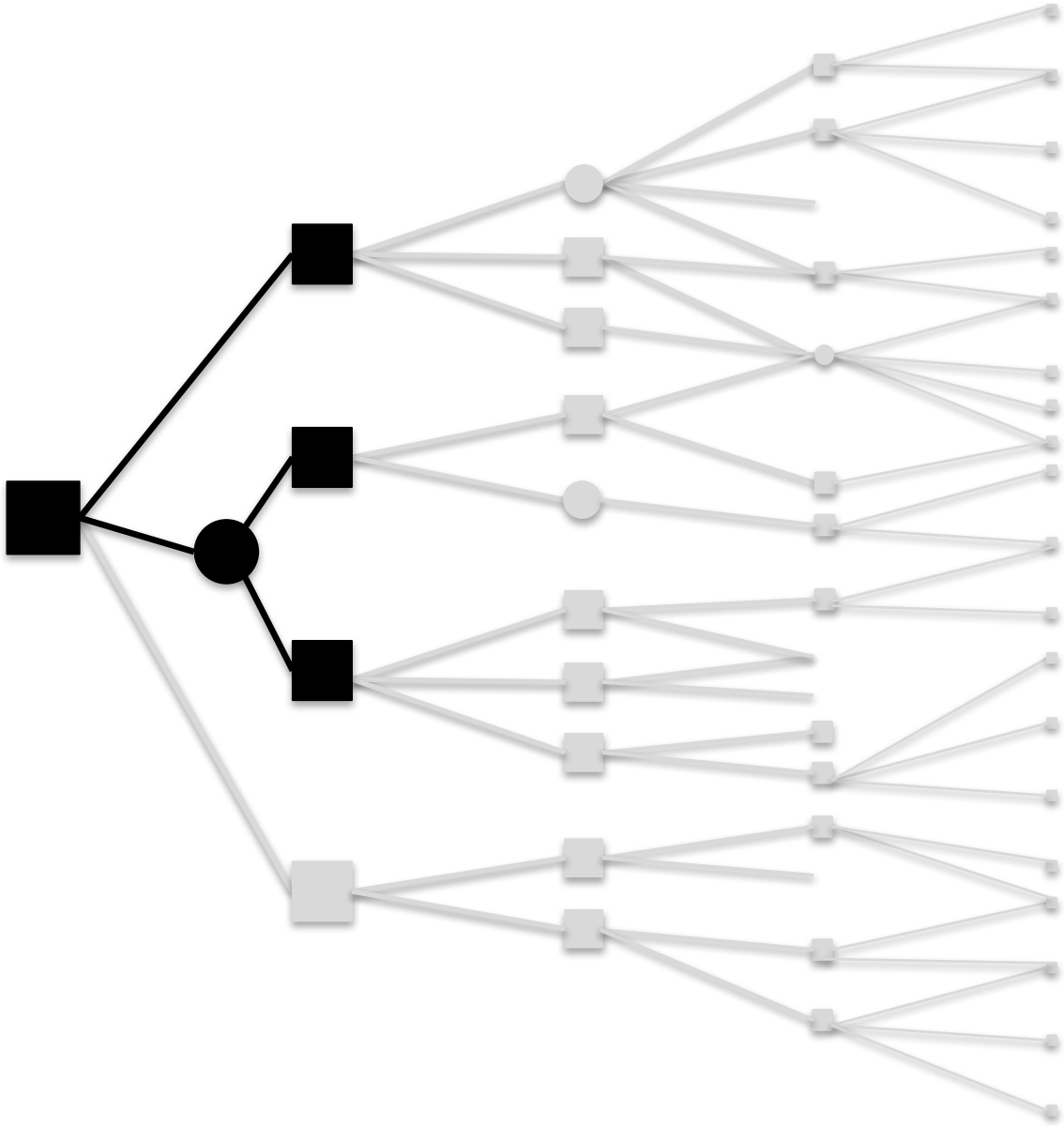
Grand World



Small World

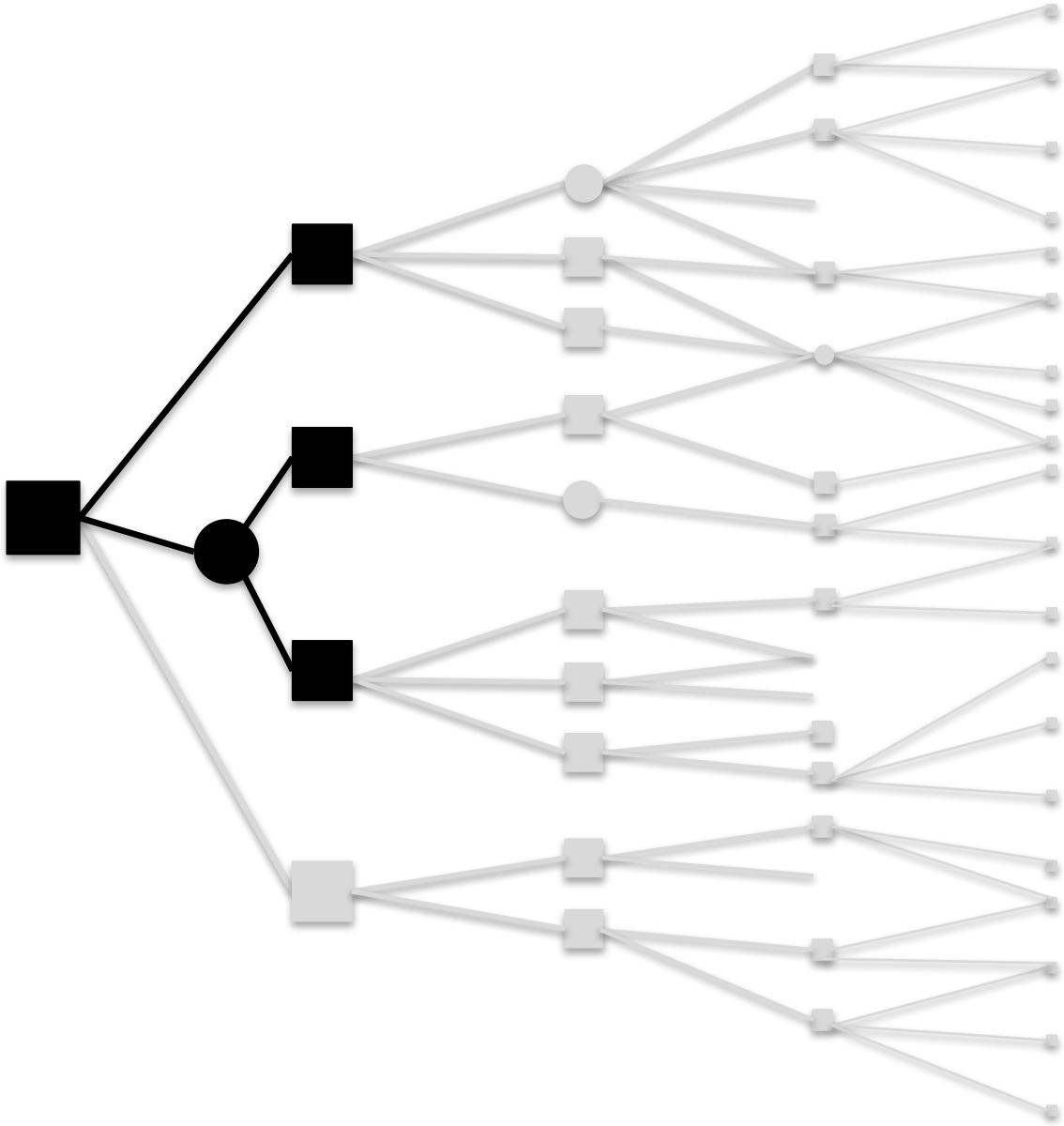
We begin with a small world, but model agents *that have neither a fixed set of acts nor settled values*.

Instead, they begin with a coarse representation of their decision situation in which they have an imperfect and provisional understanding of both the options available and the values at stake.

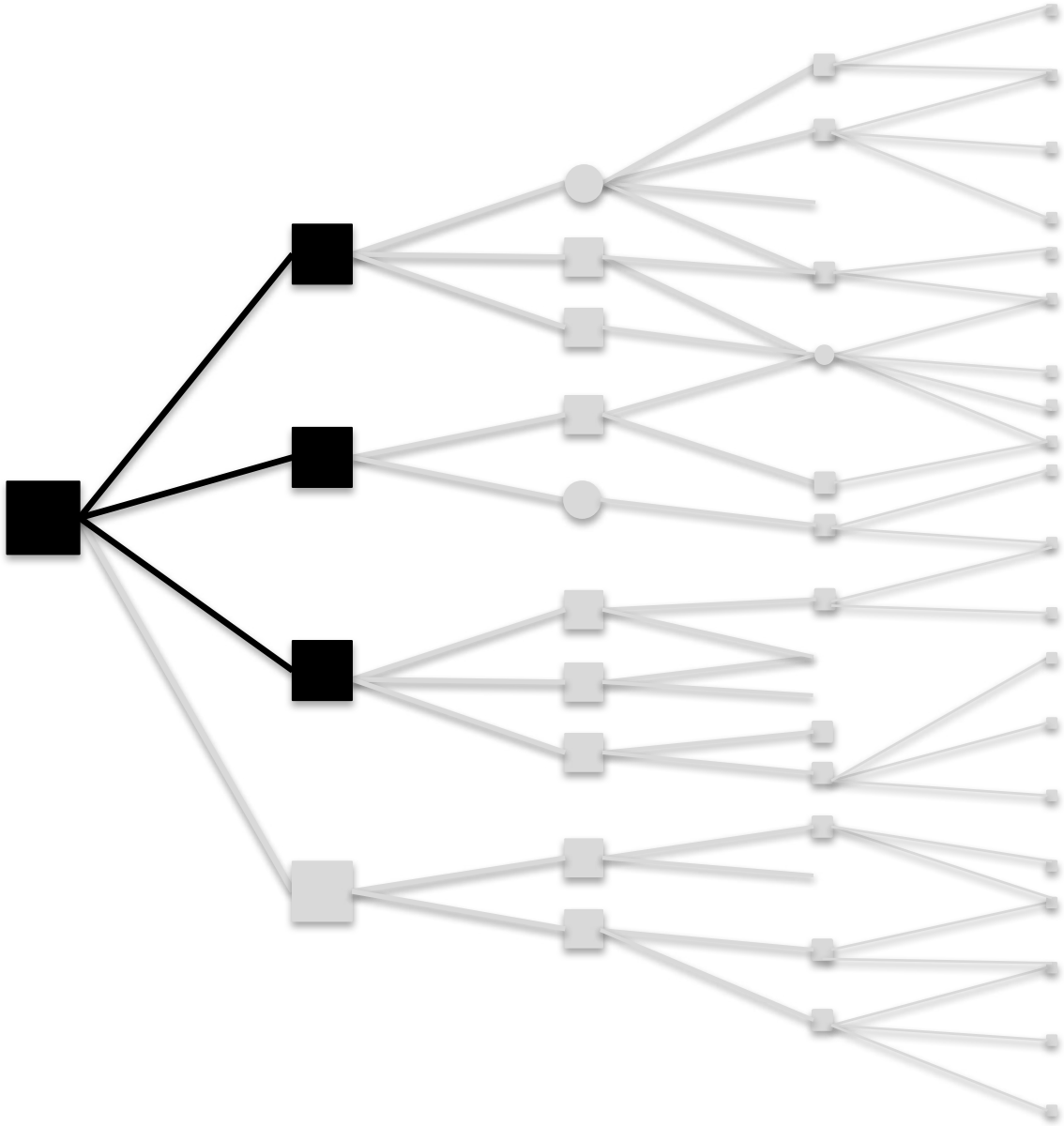


Provisional
Small World

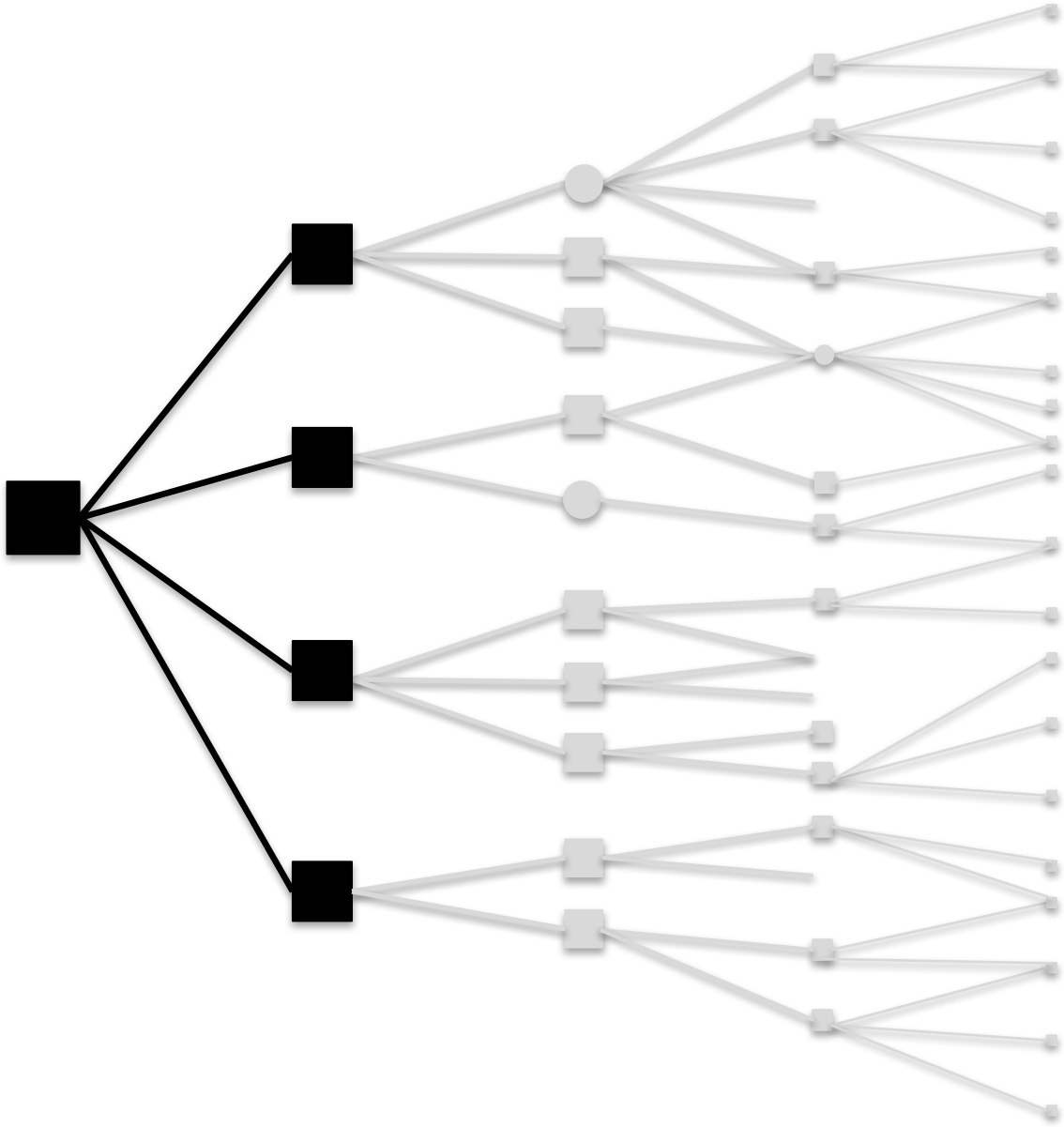
However, agents can refine their understanding of the decision problem incrementally and when the benefit of such refinement outweighs its cost.



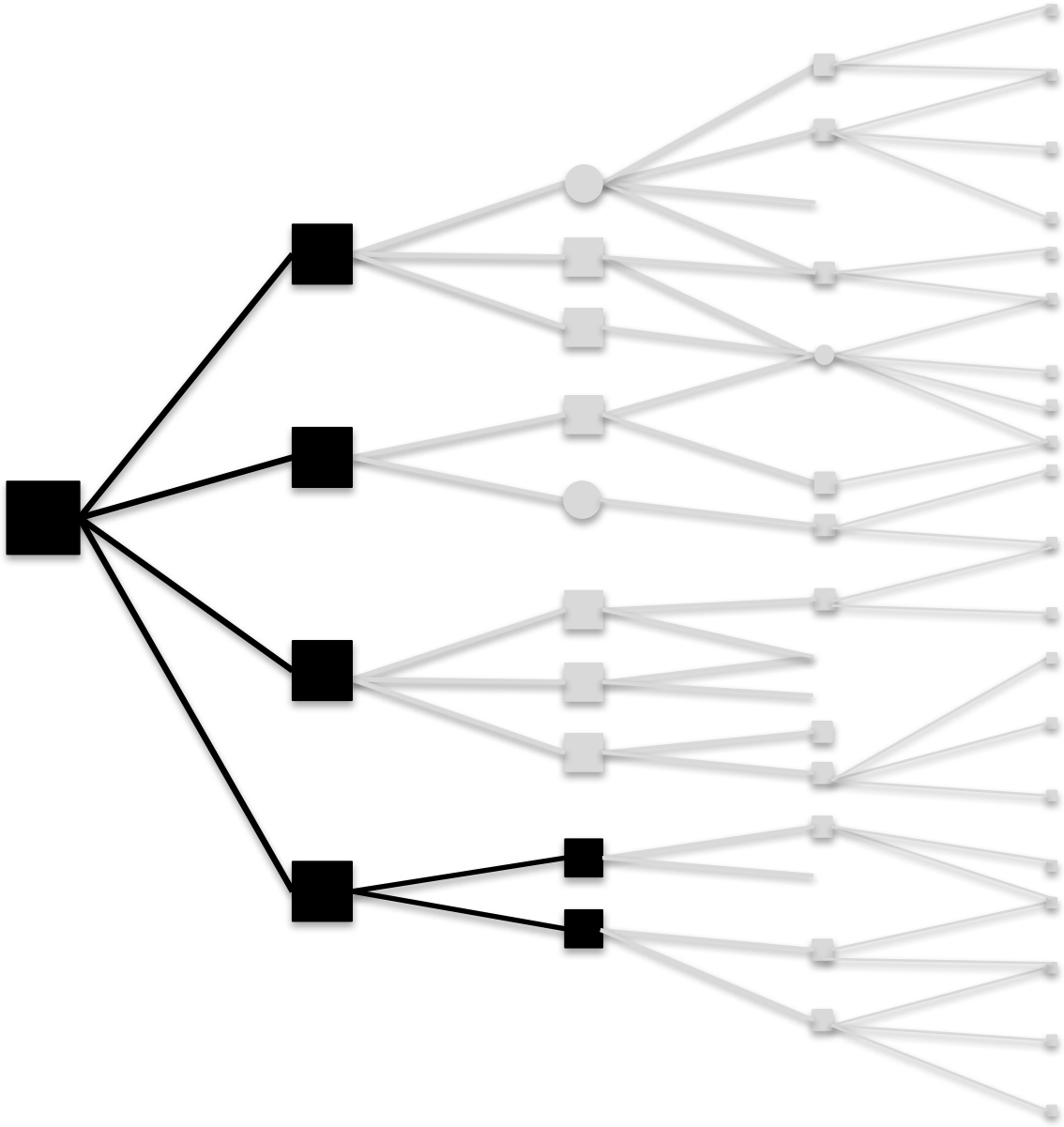
Provisional
Small World at t_0



Provisional
Small World at t_1



Provisional
Small World at t_2



Provisional
Small World at t_3

To model this, we extend the Jeffrey-Bolker framework to model uncertainty regarding our acts and utilities.

Our Results

In the **single-agent setting**:

- value-of-information theorem for value refinement
- address dilemmas of commensuration

In the **multi-agent setting**:

- transform zero-sum games into positive-sum interactions
- yields Pareto-improvements in Nash bargaining.

For this talk, we will only consider...

In the **single-agent setting**:

- value-of-information theorem for value refinement
- address dilemmas of commensuration

In the **multi-agent setting**:

- transform zero-sum games into positive-sum interactions
- yields Pareto-improvements in Nash bargaining.

For this talk, we only consider...

In the **single-agent setting**:

- value-of-information theorem for value refinement
- address dilemmas of commensuration

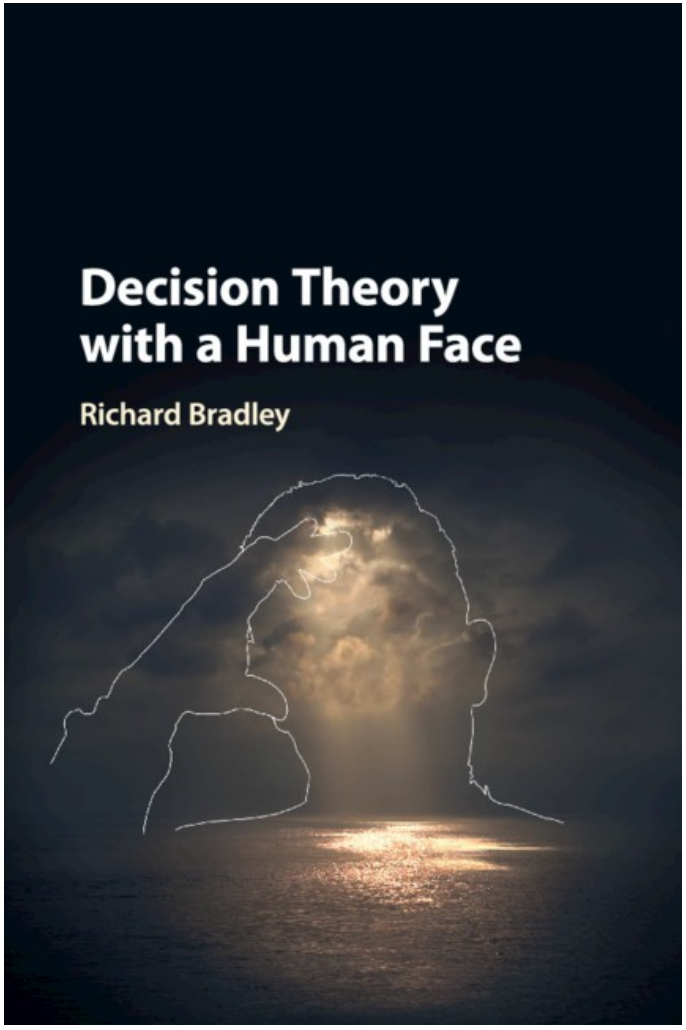
In the **multi-agent setting**:

- transform zero-sum games into positive-sum interactions
- yields Pareto-improvements in Nash bargaining.

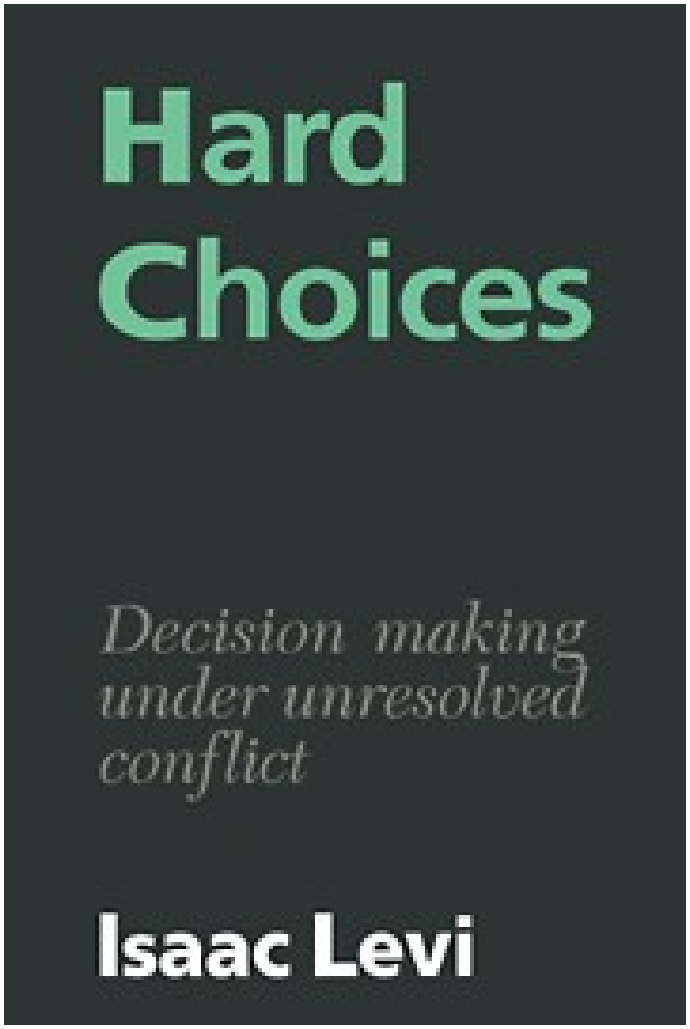
This matters for:

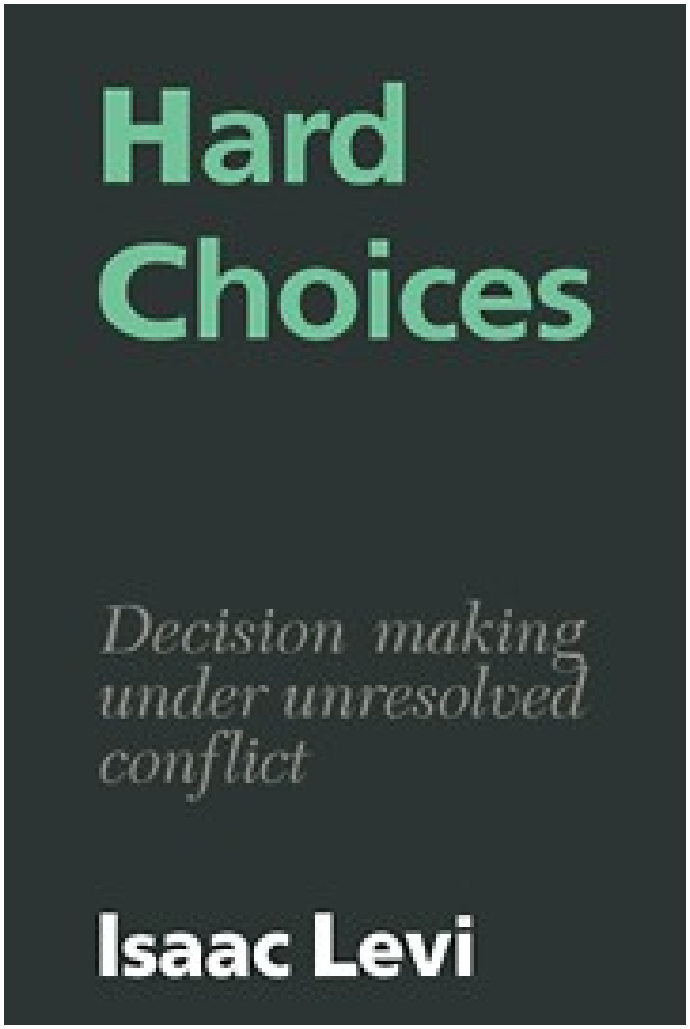
- Foundations of rational choice
- Model mis-specification in stats & ML
- Value learning in AI alignment

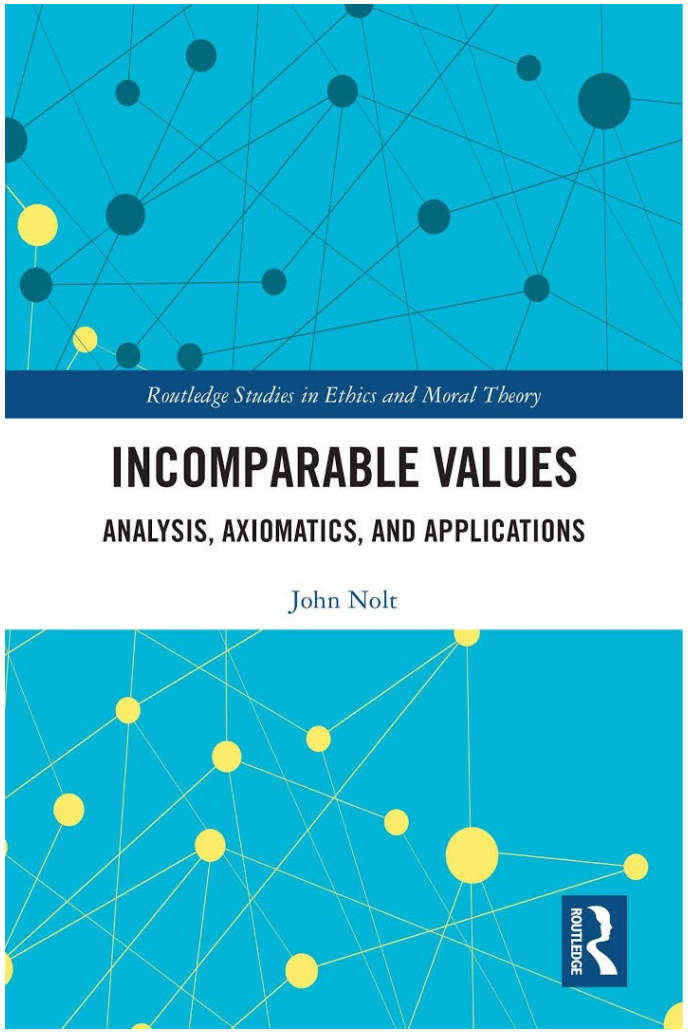
Background

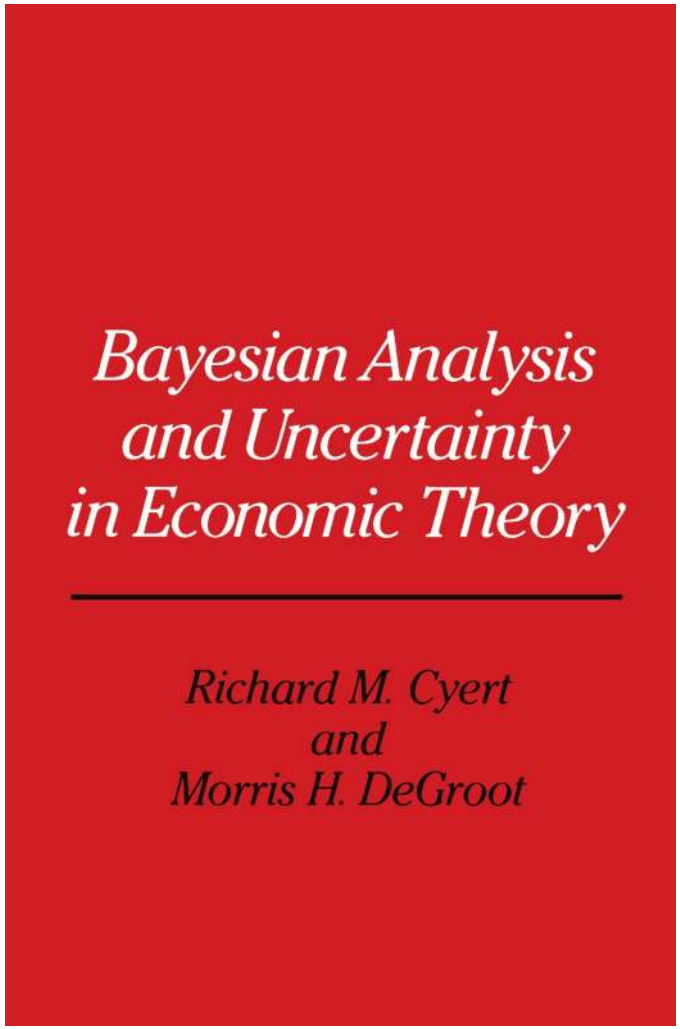

Decision Theory
with a Human Face
Richard Bradley

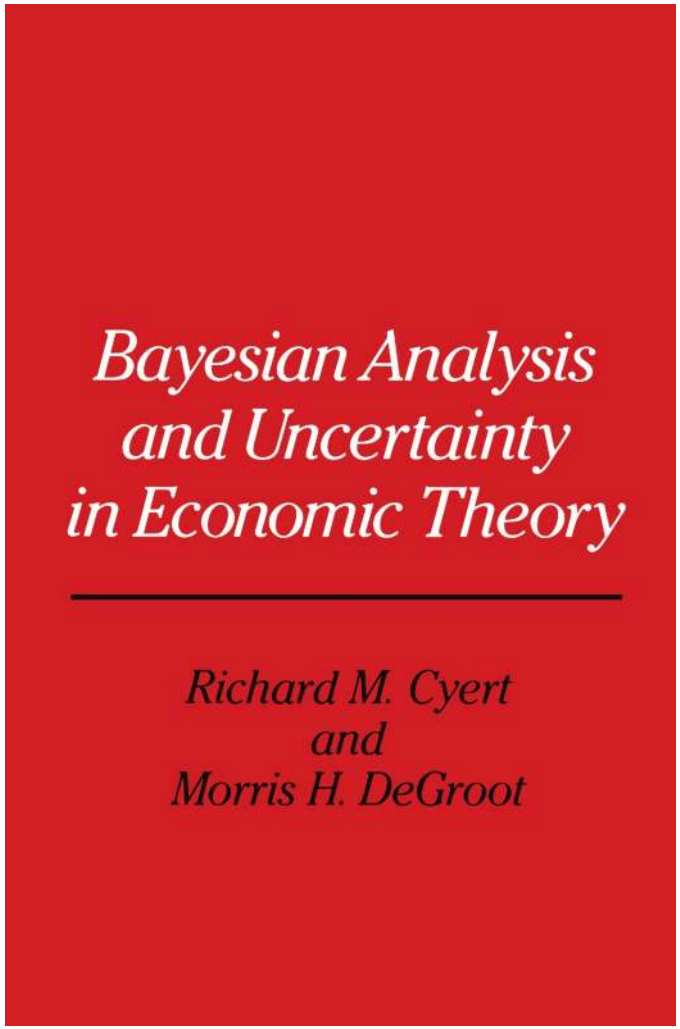

Cambridge
Elements
Decision Theory
and Philosophy
Beyond
Uncertainty
Katie Steele and
H. Orri Stefánsson


Hard
Choices
Decision making
under unresolved
conflict
Isaac Levi


Hard
Choices
Decision making under unresolved conflict
Isaac Levi


Routledge Studies in Ethics and Moral Theory
INCOMPARABLE VALUES
ANALYSIS, AXIOMATICS, AND APPLICATIONS
John Nolt
ROUTLEDGE


Bayesian Analysis
and Uncertainty
in Economic Theory
Richard M. Cyert
and
Morris H. DeGroot



arXiv:cs/0603020v1 [cs.LO] 5 Mar 2006

Reasoning About Knowledge of Unawareness*

Joseph Y. Halpern
Computer Science Department
Cornell University, U.S.A.
e-mail: halpern@cs.cornell.edu

Leandro Chaves Rêgo
School of Electrical and Computer Engineering
Cornell University, U.S.A.
e-mail: lcr26@cornell.edu

September 24, 2018

Abstract

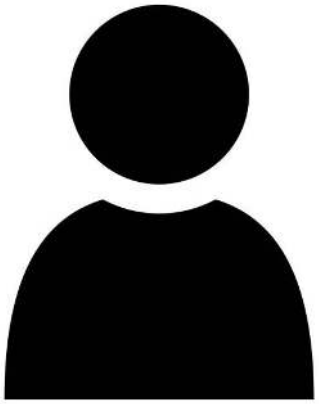
Awareness has been shown to be a useful addition to standard epistemic logic for many applications. However, standard propositional logics for knowledge and awareness cannot express the fact that an agent knows that there are facts of which he is unaware without there being an explicit fact that the agent knows he is unaware of. We propose a logic for reasoning about knowledge of unawareness, by extending Fagin and Halpern's *Logic of General Awareness*. The logic allows quantification over variables, so that there is a formula in the language that can express the fact that "an agent explicitly knows that there exists a fact of which he is unaware". Moreover, that formula can be true without the agent explicitly knowing that he is unaware of any particular formula. We provide a sound and complete axiomatization of the logic, using standard axioms from the literature to capture the quantification operator. Finally, we show that the validity problem for the logic is recursively enumerable, but not decidable.

*This work was supported in part by NSF under grants CTC-0208535, ITR-0325453, and IIS-0534064, by ONR under grants N00014-00-1-03-41 and N00014-01-10-511, and by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the ONR under grant N00014-01-1-0795. The second author was also supported in part by a scholarship from the Brazilian Government through the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

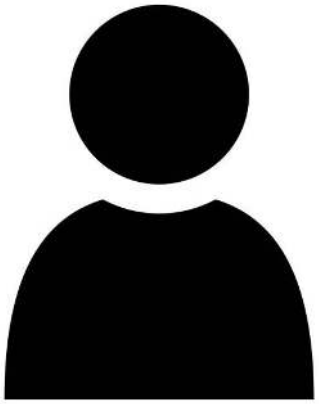
A Fruitful Vignette

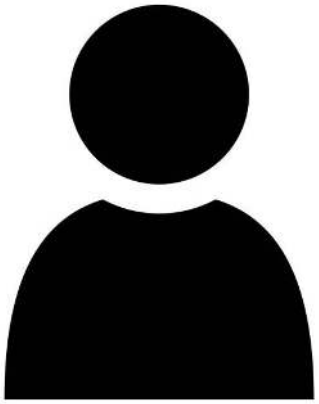
Consider two agents...

Consider two agents...



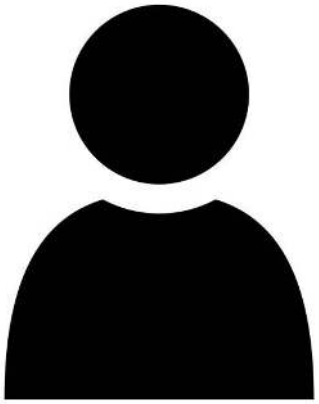
Consider two agents...

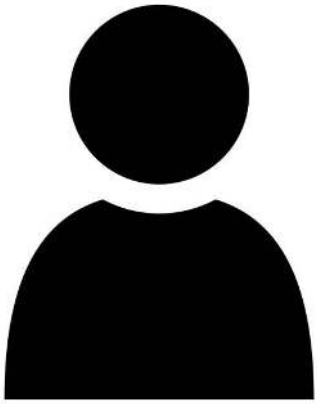




Consider two agents...

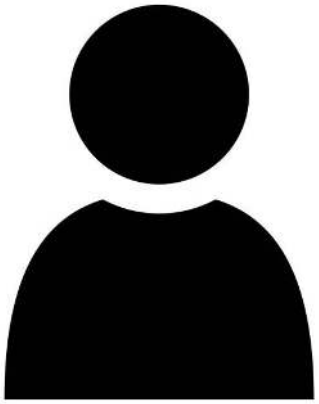
the *Arborist*



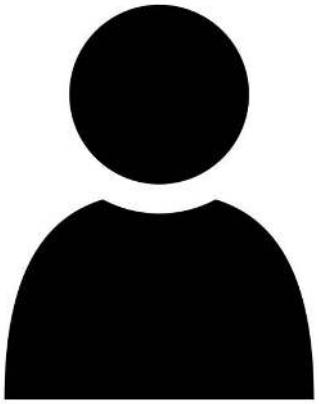


Consider two agents...

the *Arborist*

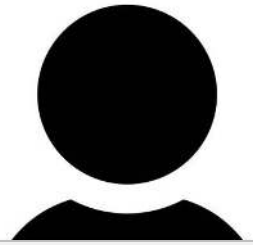


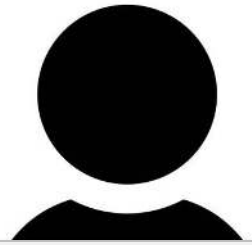
the *Baker*.



Consider two agents...

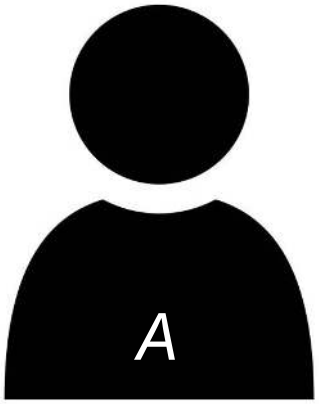
the *Arborist* the *Baker*.

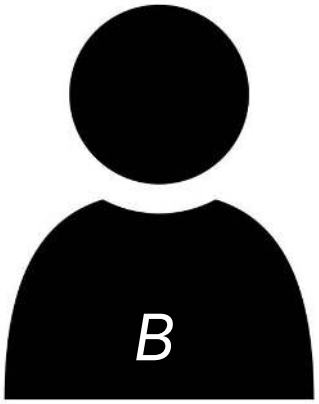




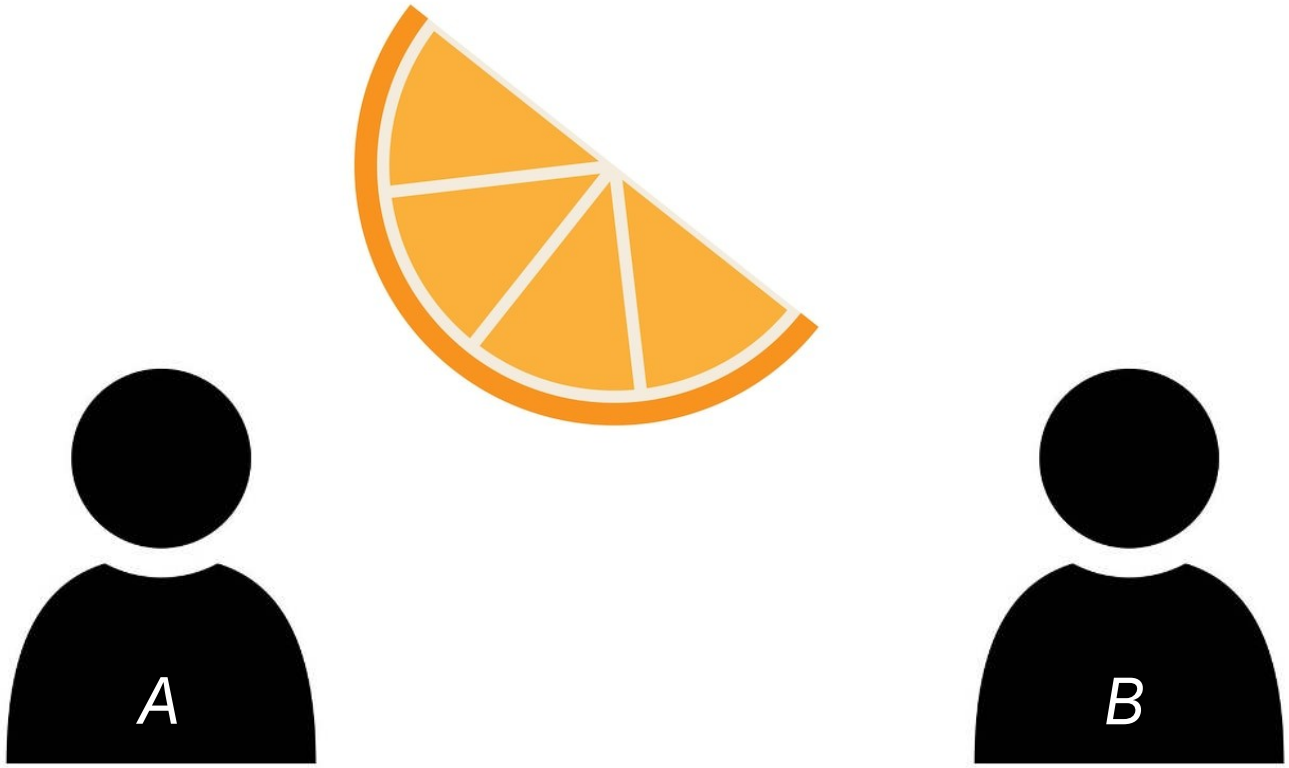
Roger Fisher & William Ury, *Getting to Yes* (1981)

Henceforth:

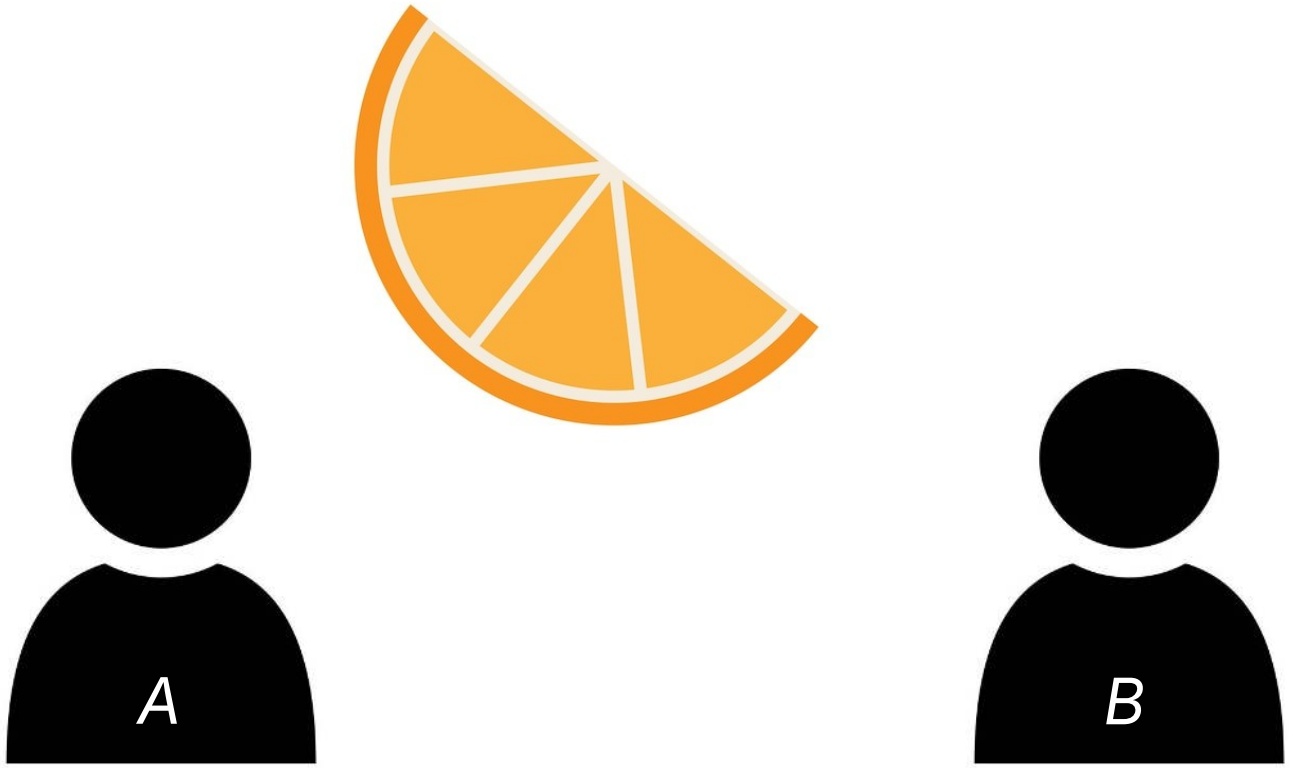




They run into an orange...

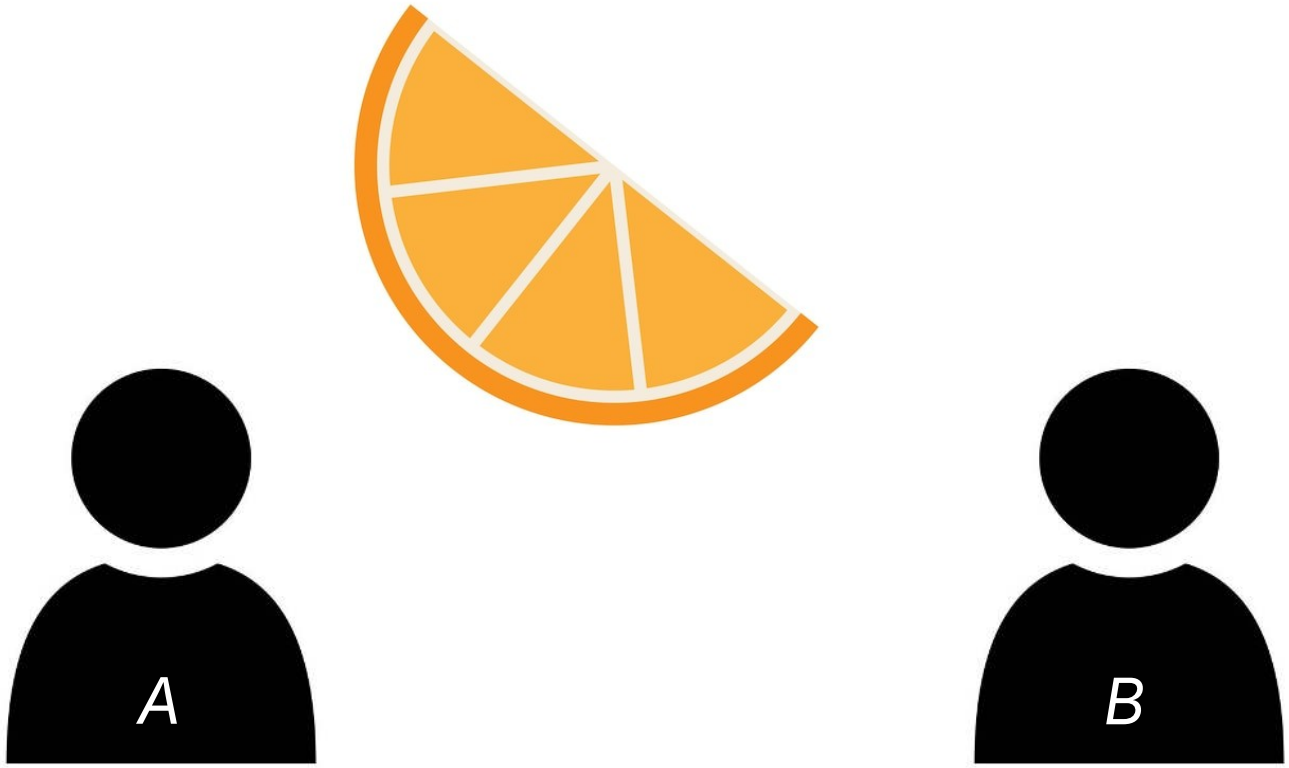


How should they split this resource?

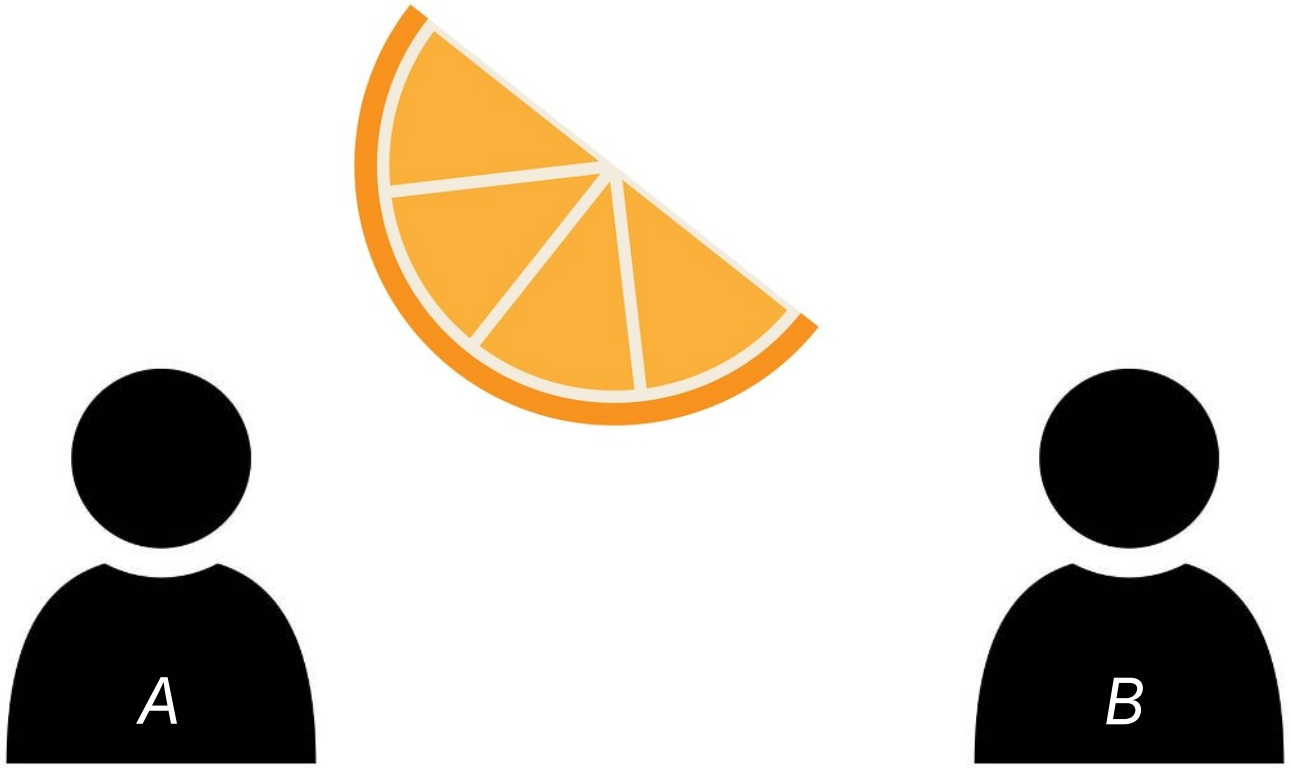


Initially:

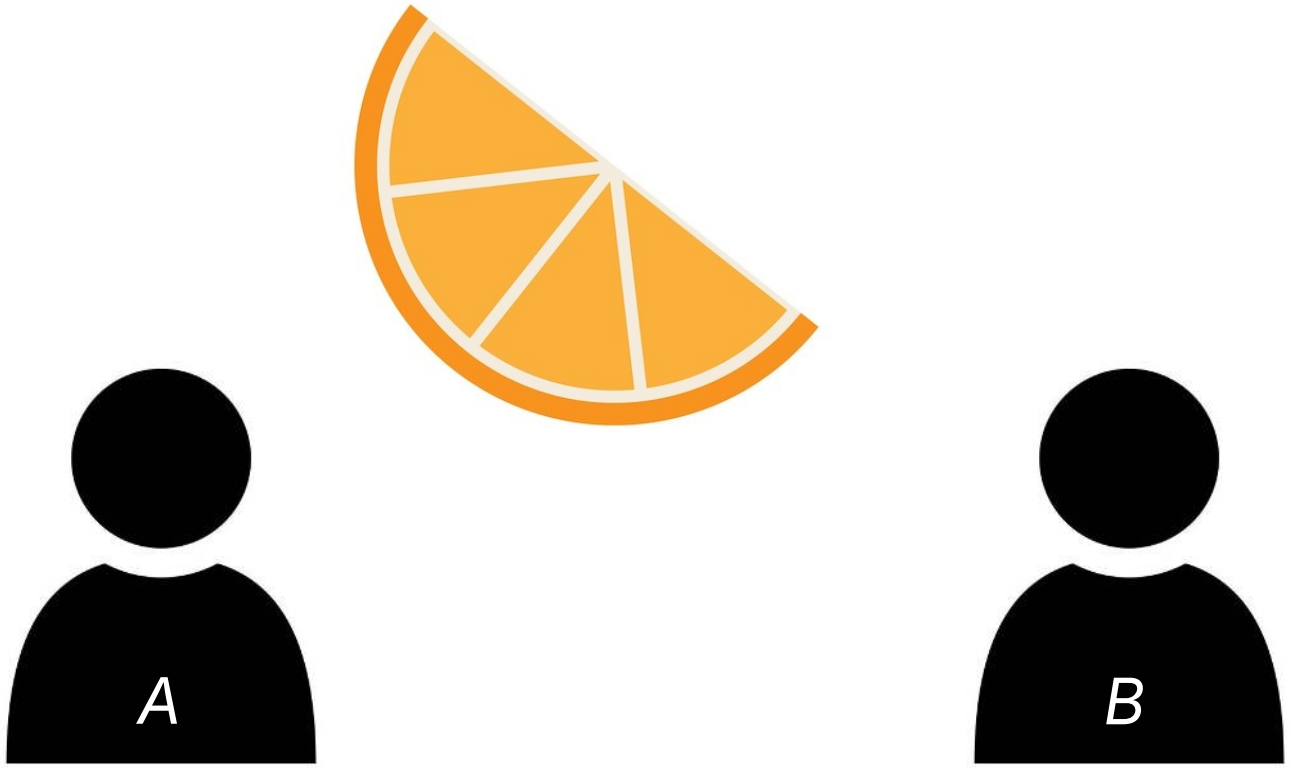
{Split the resource, Fight over the resource}



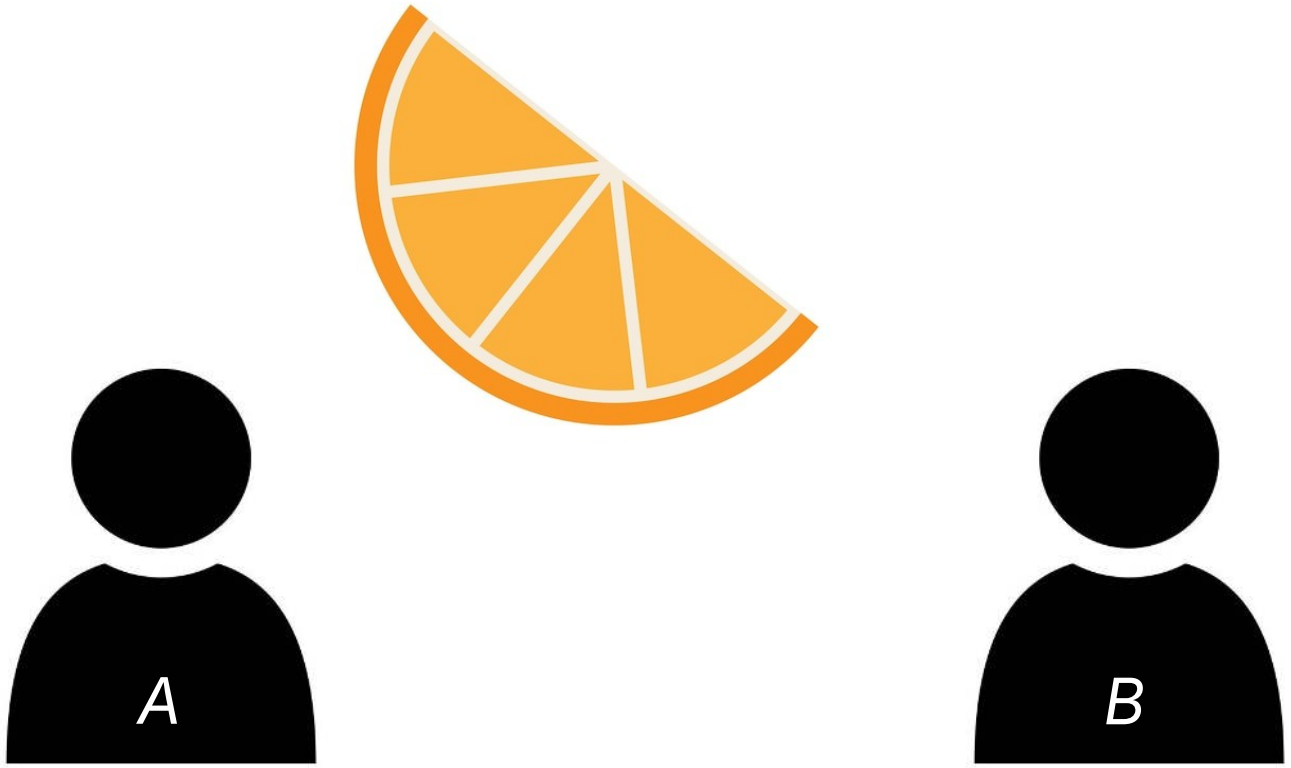
They realize:



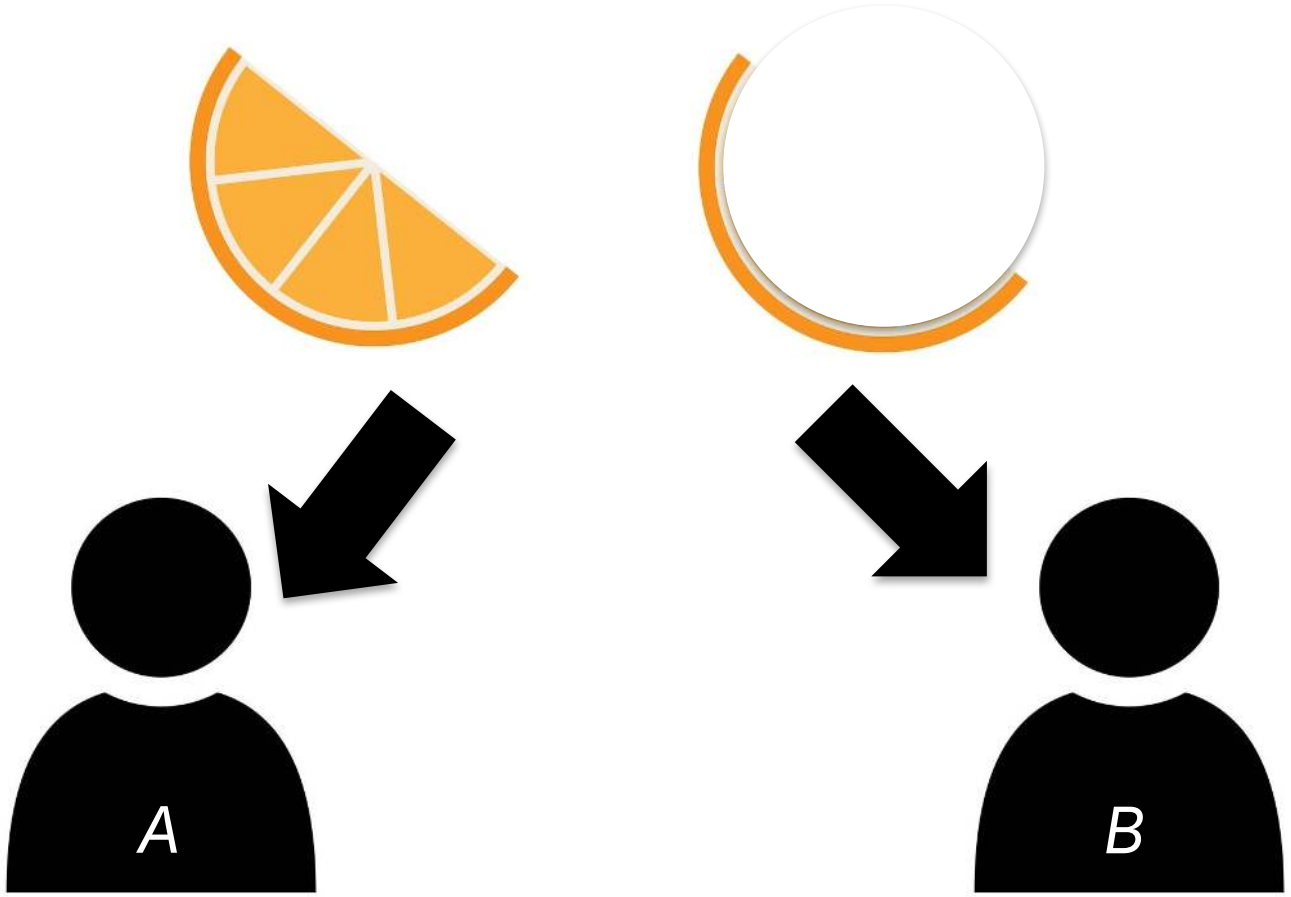
They realize:
the *Arborist* prefers the seeds (for propagating);



They realize:
the *Arborist* prefers the seeds (for propagating);
and the *Baker* prefers the rind (for baking)?



Then they might step back, reflect, and refine their options sets to include more acts, including this one.



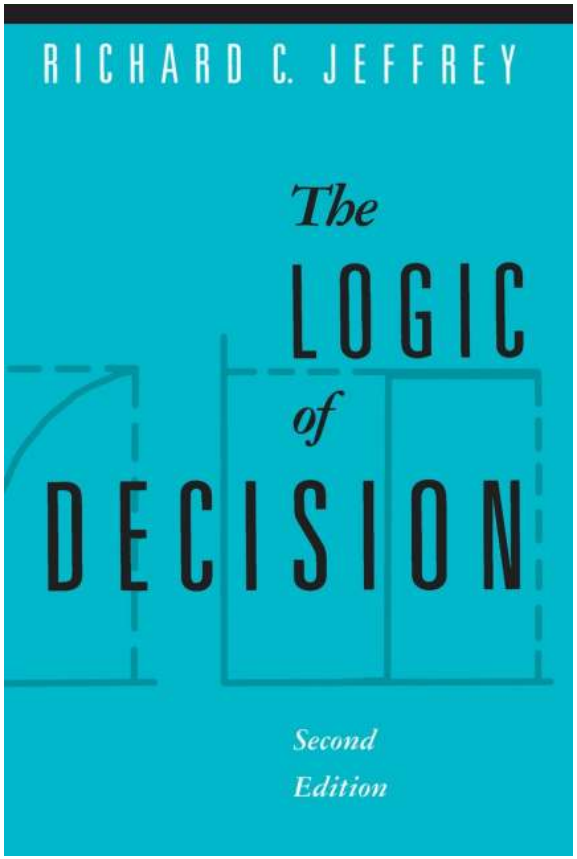
Question 1. How should we model this sort of refinement of agents' options and values?

Question 2. When is it rational for agents to refine their options and values, as opposed to treating their current model as sufficient?

Question 3. What, if any, are the interesting consequences of this sort of refinement for strategic interactions?

Our Model

The Jeffrey-Bolker decision framework



Here, acts, states, and consequences are the same type of object: propositions in an atomless Boolean algebra.

Jeffrey-Bolker is apt her because the agent’s algebra can be indefinitely refined

Modeling Refinement

First, consider an omniscient benchmark:

$\mathfrak{A}$ complete, atomless Boolean algebra

$\mathbb{A}$ includes all maximally fine-grained acts

This agent has already considered every relevant factor to arbitrary precision and settled her valuation of every possible act.

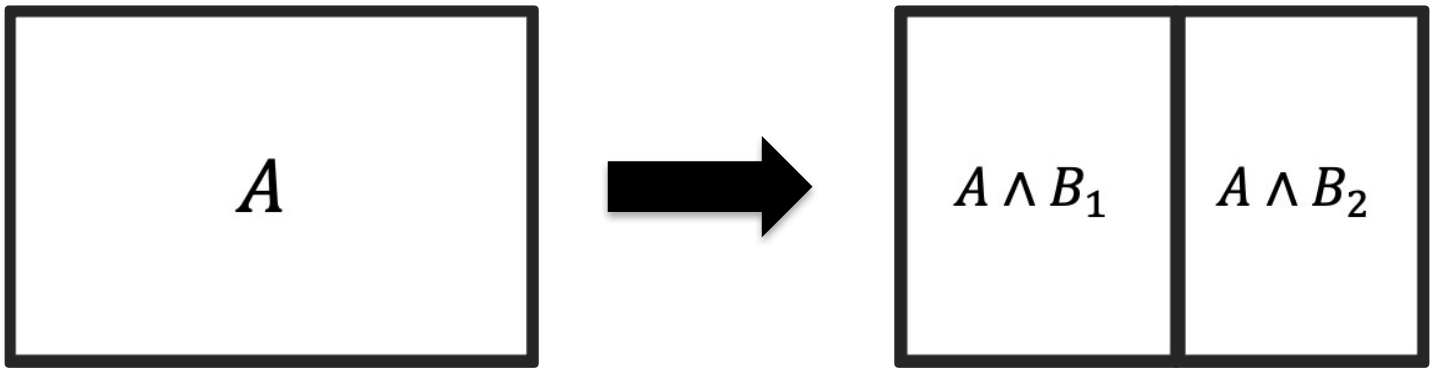
$P: \mathfrak{A} \to [0,1]$ and $U: \mathfrak{A}' \to \mathbb{R}$ are fully specified

In contrast, the agent we are interested in modeling has not undergone such arbitrary refinement.

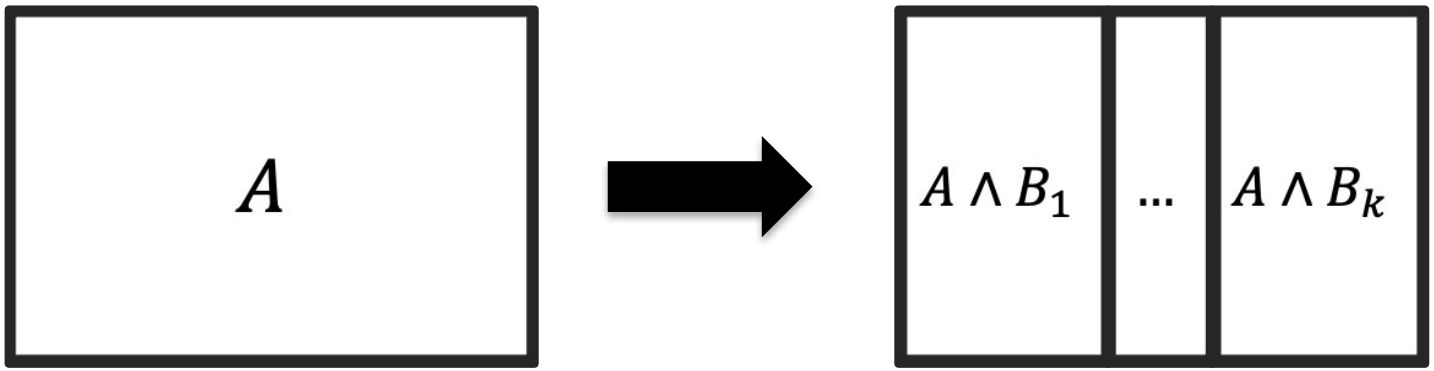
This agent has not considered every factor, and her act are restricted to those options she has considered.

What does refinement look like for such an agent?

Binary Refinement



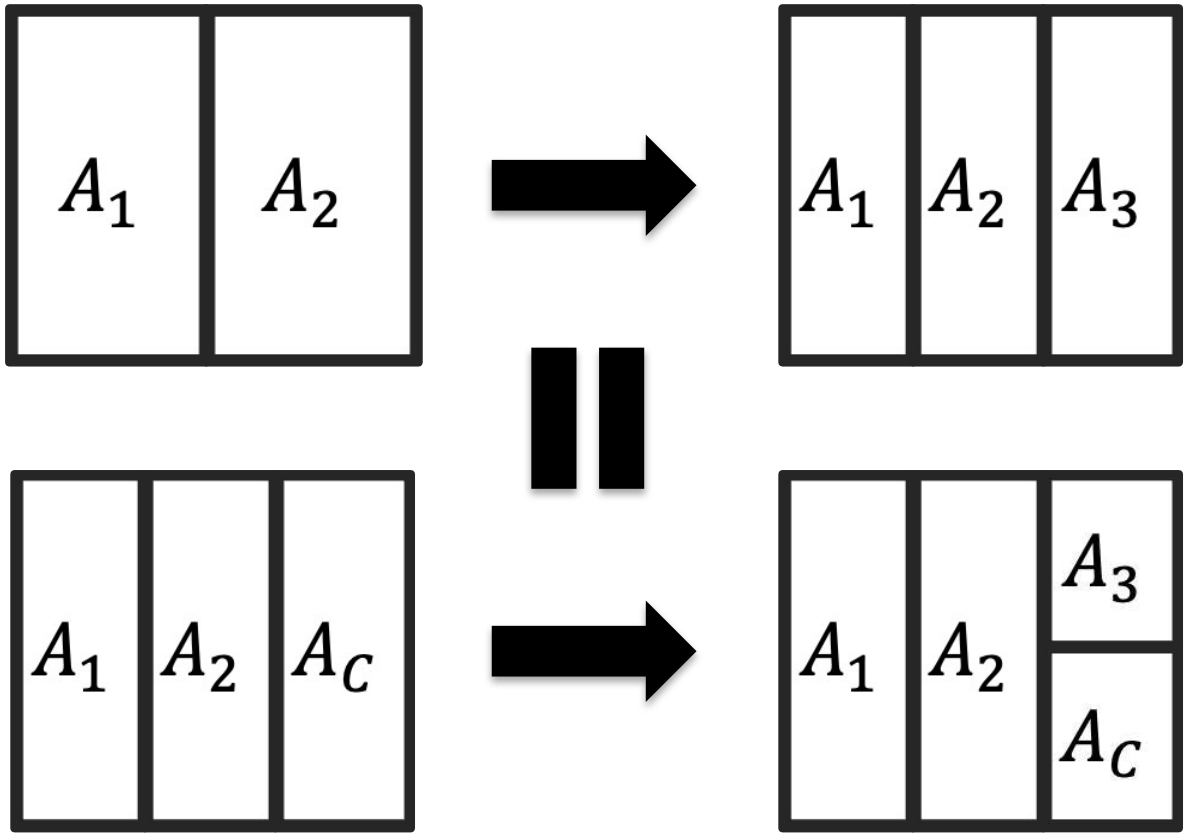
K-ary Refinement



N.B.: Any *k*-ary refinement can be achieved through sequential binary refinements; so we focus on the binary case throughout.

Also, if an agent is equipped with a catch-all option, A_c, then **expansion** can be modeled as a **special case of refinement**—refinement of the catch-all.

Expansion via refinement with a catch-all



Value refinement then involves two distinct stages:

i. *Ex ante* uncertainty about options and values;

ii. *Ex post* realization of options and values.

So, we need to model the agent's *attitudes toward possible refinements*.

To do this we equip them with a probability space.

Each $\omega \in \Omega$ represents one possible realization of refinement

Define the refinement outcome mapping

$$\xi^A: \Omega \to \mathbb{R}^2 \times (0,1)^2$$

by

$$\xi^A(\omega) := (u_1(\omega), u_2(\omega), p_1(\omega), p_2(\omega))$$

where

$$u_1 = U_1(A \wedge B_1)$$

$$u_2 = U_1(A \wedge B_2)$$

$$p_1 = P(B_1)$$

$$p_2 = P(B_2)$$

$$q = \frac{p_1}{p_1 + p_2} = P(B_1 \mid A)$$

Note that it's possible that

$$p_1(\omega) + p_2(\omega) \neq P(A)$$

and

$$P_0(\mathrm{A})/P_0(\neg\mathrm{A}) \neq P_1(\mathrm{A})/P_1(\neg\mathrm{A})$$

i.e., reverse rigid extension is possible but not assumed

The agent's beliefs about the outcomes of the refinement of act A are captured by the distribution:

$$\mu_A$$

What constraints should a rational agent’s credence and utilities have w.r.t to possible refinements?

Our proposal: the agent's current valuation of a proposition should equal her expected future valuation after refinement.

Definition (Refinement Reflection Principle). For $A \in \mathbb{A}_0$ and binary refinement $\mathcal{R}_A = \{A \wedge B_1, A \wedge B_2\}$, a rational agent's valuation obeys:

$$U_0(A) = \mathbb{E}_{\mu_A}[U_1(A)] = \mathbb{E}_{\mu_A}[qu_1 + (1-q)u_2]$$

Where $(u_1, u_2, q) \sim \mu_A$.

Traditional reflection principles:

$$P_t(A) = \mathbb{E}_t[P_{t+s}(A)]$$

Our refinement reflection principles:

$$U_t(A) = \mathbb{E}_t[U_{t+s}(A)]$$

N.B.: We do not take reflection to be a universal principle of rationality; it will sometimes be violated

(e.g., in cases of act-state dependence, finitely but not countably additive probabilities, imprecise probabilities, etc.)

Our second condition ensures genuine uncertainty about refinement outcomes.

Definition (Refinement Uncertainty). An agent has *refinement uncertainty* regarding an act A if their credences μ with respect to the outcomes of refinement satisfies:

$$\mu(u_1 \neq u_2) > 0.$$

N.B.: Similarly, it is not incoherent for an agent to be dogmatically opposed to the possibility of learning.

That said we expect these conditions to obtain in a broad range of cases; and these will be the cases of interest for us.

Together, these conditions ensure positive probability of discovering strict improvements.

Value of Value Refinment

Question. When agents will value refinement?

Answer. When refinement is *cost-free*, and the agent's credences satisfies RRP, and refinement uncertainty.

$V_0 :=$ value of best responses pre-refinement

$V_1 :=$ the value of best response post-refinement

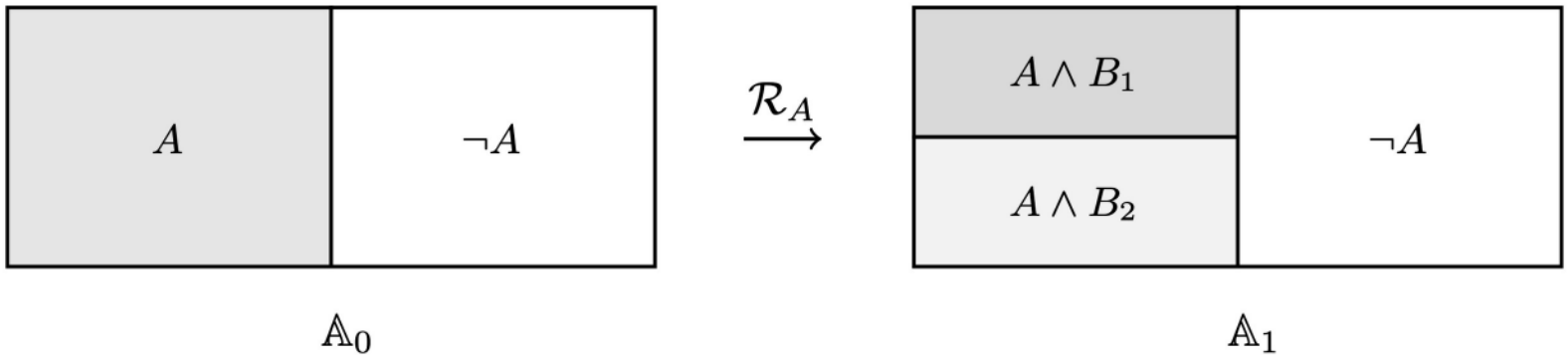
Theorem (The Value of Value Refinement). Let an agent's credence satisfy RRP and refinement uncertainty, then:

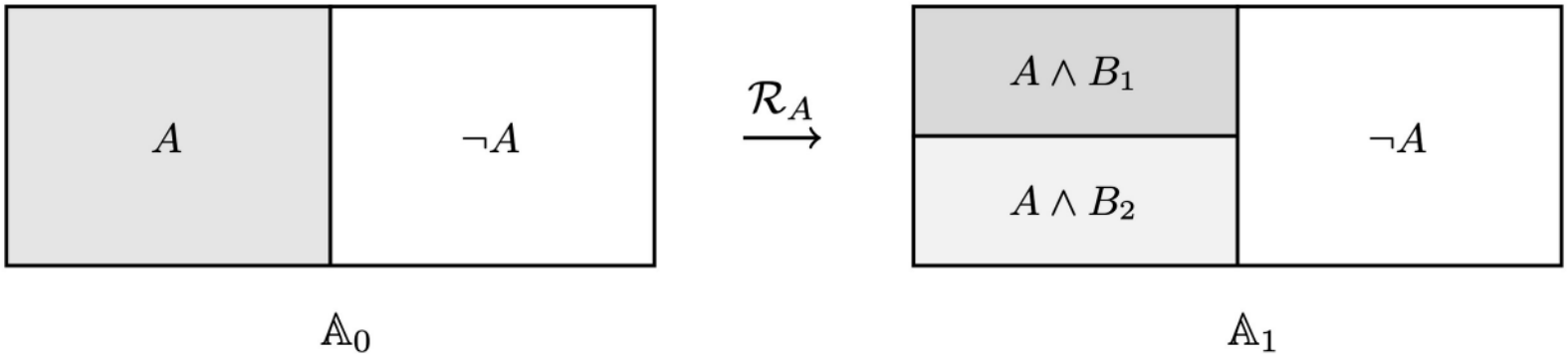
$$\mathbb{E}_{\mu_A}[V_1] > V_0$$

Intuitively, refinement transforms commitment to an average of a coarse-grained bundle into a choice among its fine-grained components.

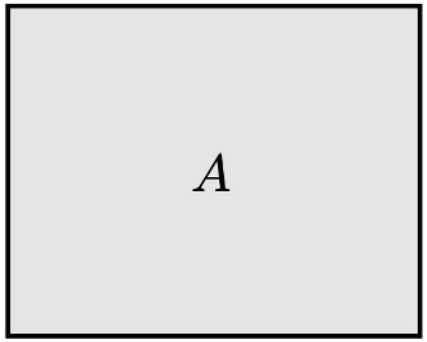


$\mathbb{A}_0$

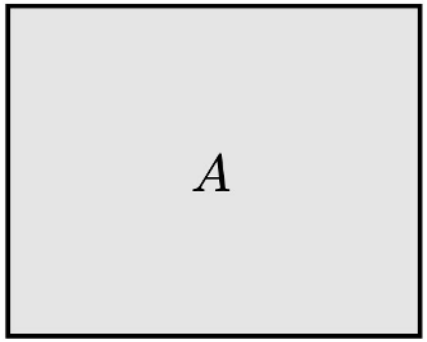

A
$\neg A$
$\mathcal{R}_A$
$A \wedge B_1$
$A \wedge B_2$
$\neg A$
$\mathbb{A}_0$
$\mathbb{A}_1$



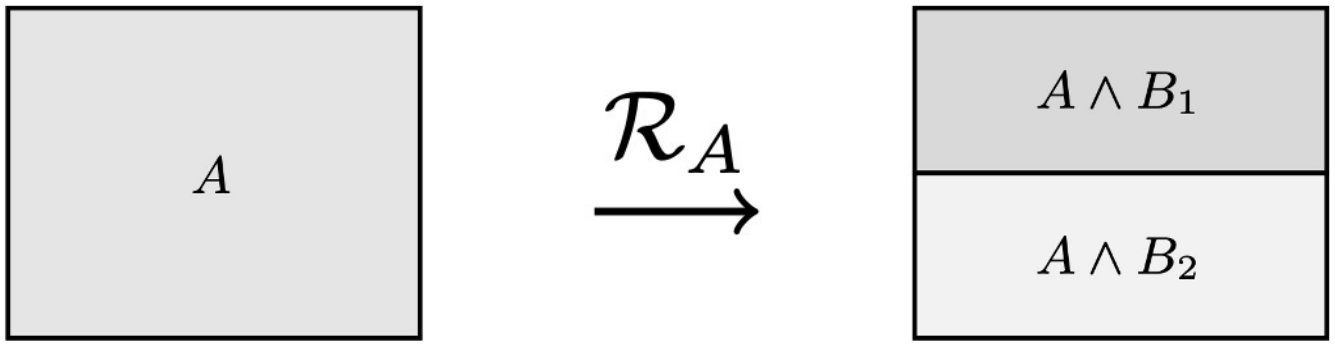
- Let's focus on A—the act to be refined


A

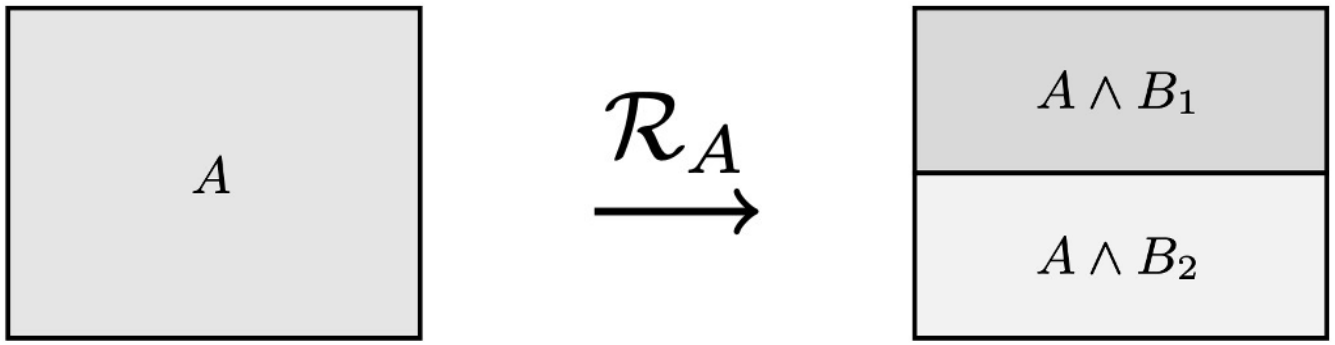
$$U(A) = \mathbb{E}_{\mu_A}\left[qu_1 + (1-q)u_2\right]$$



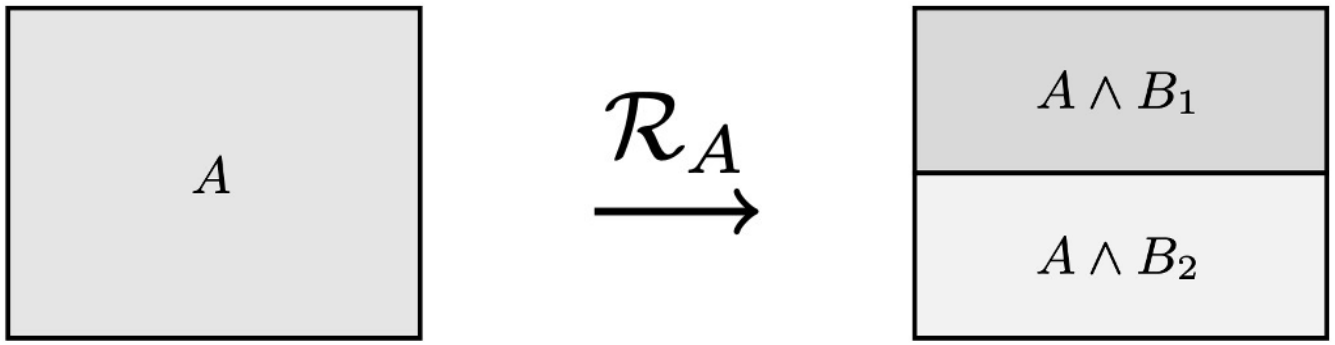
$$U(A) = \mathbb{E}_{\mu_A}\left[qu_1 + (1-q)u_2\right]$$



$$U(A) = \mathbb{E}_{\mu_A}\left[qu_1 + (1-q)u_2\right] \qquad \mathbb{E}_{\mu_A}\left[\max\{u_1, u_2\}\right]$$



$$U(A) = \mathbb{E}_{\mu_A}\left[qu_1 + (1-q)u_2\right] \qquad \mathbb{E}_{\mu_A}\left[\max\{u_1, u_2\}\right]$$

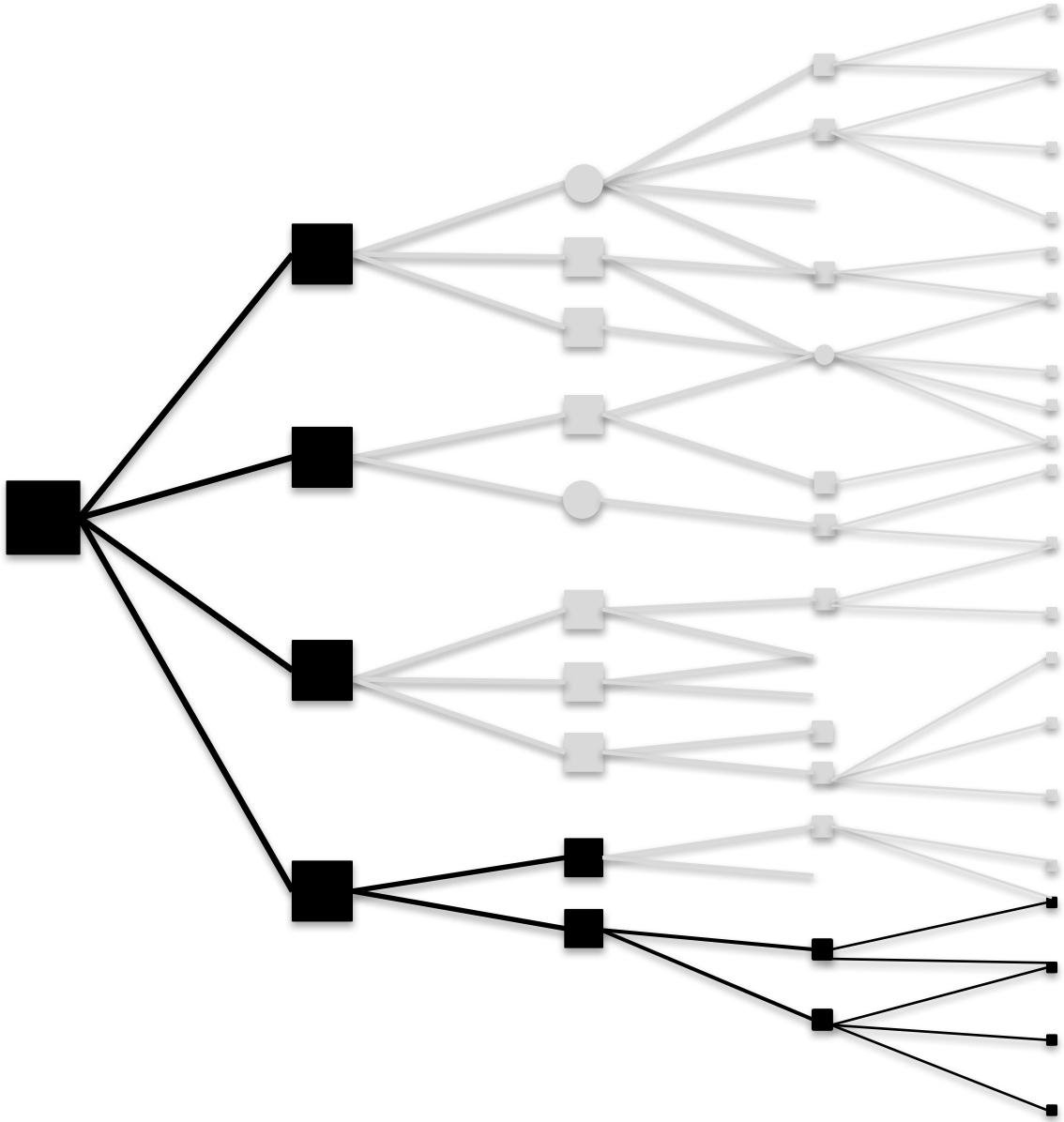


This yields a Ramsey- or Good-style value of information theorem, but for refinement.

In our paper, we also characterize optimal stopping policies for refinement *with costs*

Taking stock

We’ve filled the lacuna in Savages picture of how decision theory can be applied to rational agents who are uncertain about their options and values



Rational choice for such agents often has a two-stage character: *refine, then maximize*

We showed how refinement can produce **gains in individual decision-making**...

You can read our paper to see how refinement yields:

- The possibility of **dilemma resolution**
- Turning **zero-sum games** to positive-sum
- Pareto improvements in **Nash bargaining.**

Thank you.

Link to the paper on `arXiv.org`

Institute for Complex Social Dynamics

Carnegie Mellon University

Center for AI Ethics and Policy

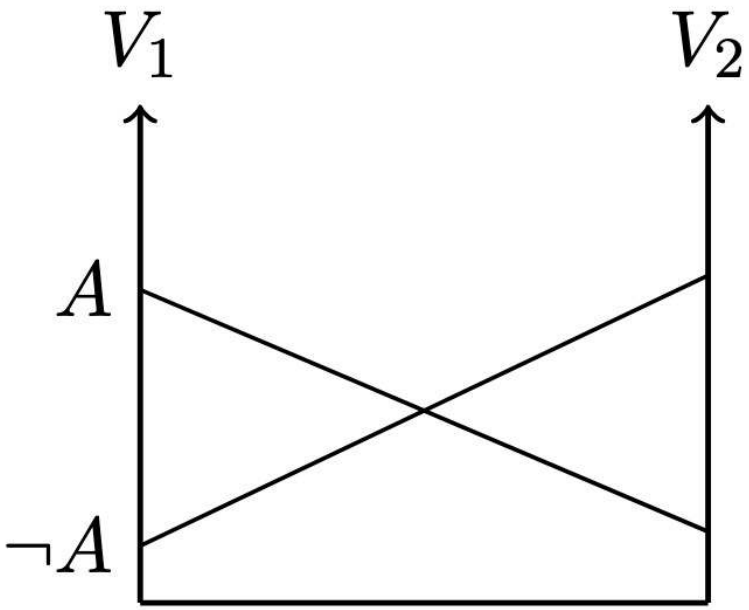
Value Commensuration

Consider the following dilemma.

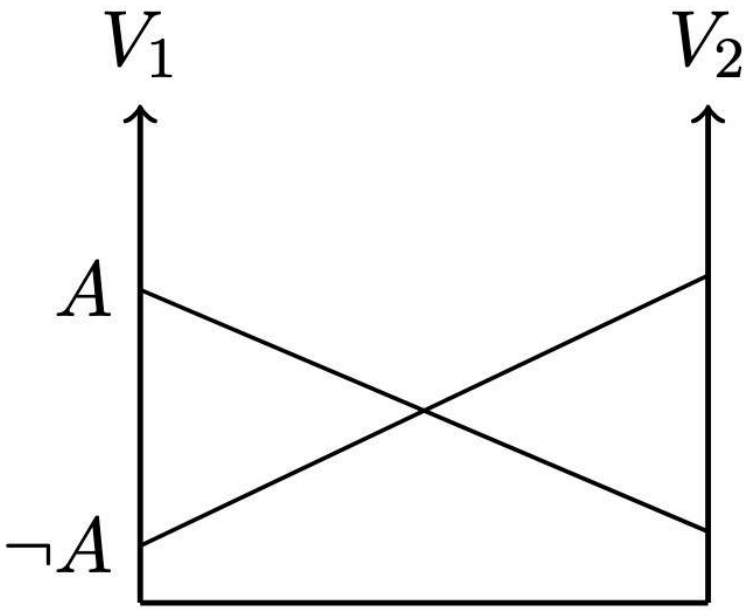
An agent has two possible acts A and $\neg$A

And two dimensions of value V_1 and V_2 within which she can evaluate her acts

It turns out that each value favors one act



It turns out that each value favors one act



But she doesn’t know how to commensurate the values

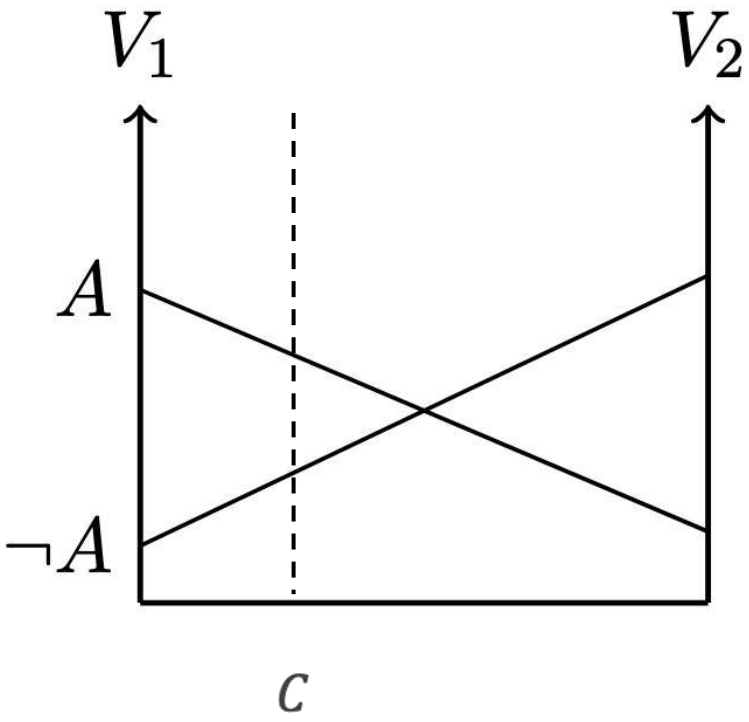
Examples.

Truth vs Justice

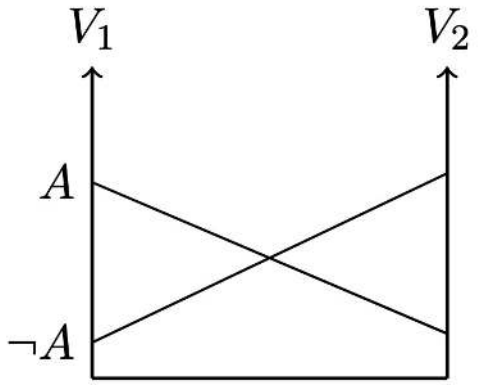
Growth vs Safety

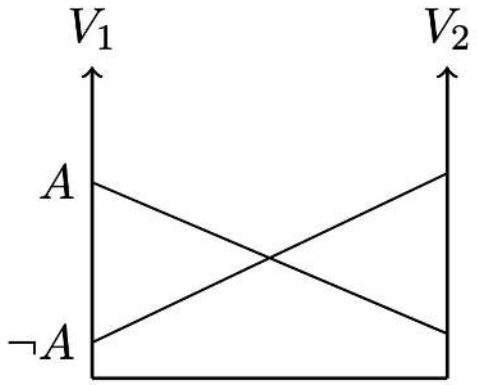
Schmeckles, Flurbos, and Blemflarcks vs USD

To commensurate them, she will have to create an aggregate value function V and determine the weight C

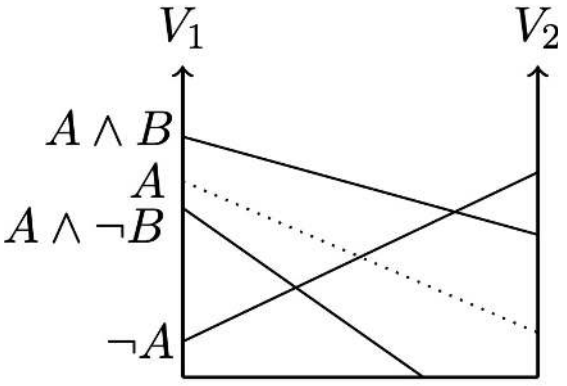


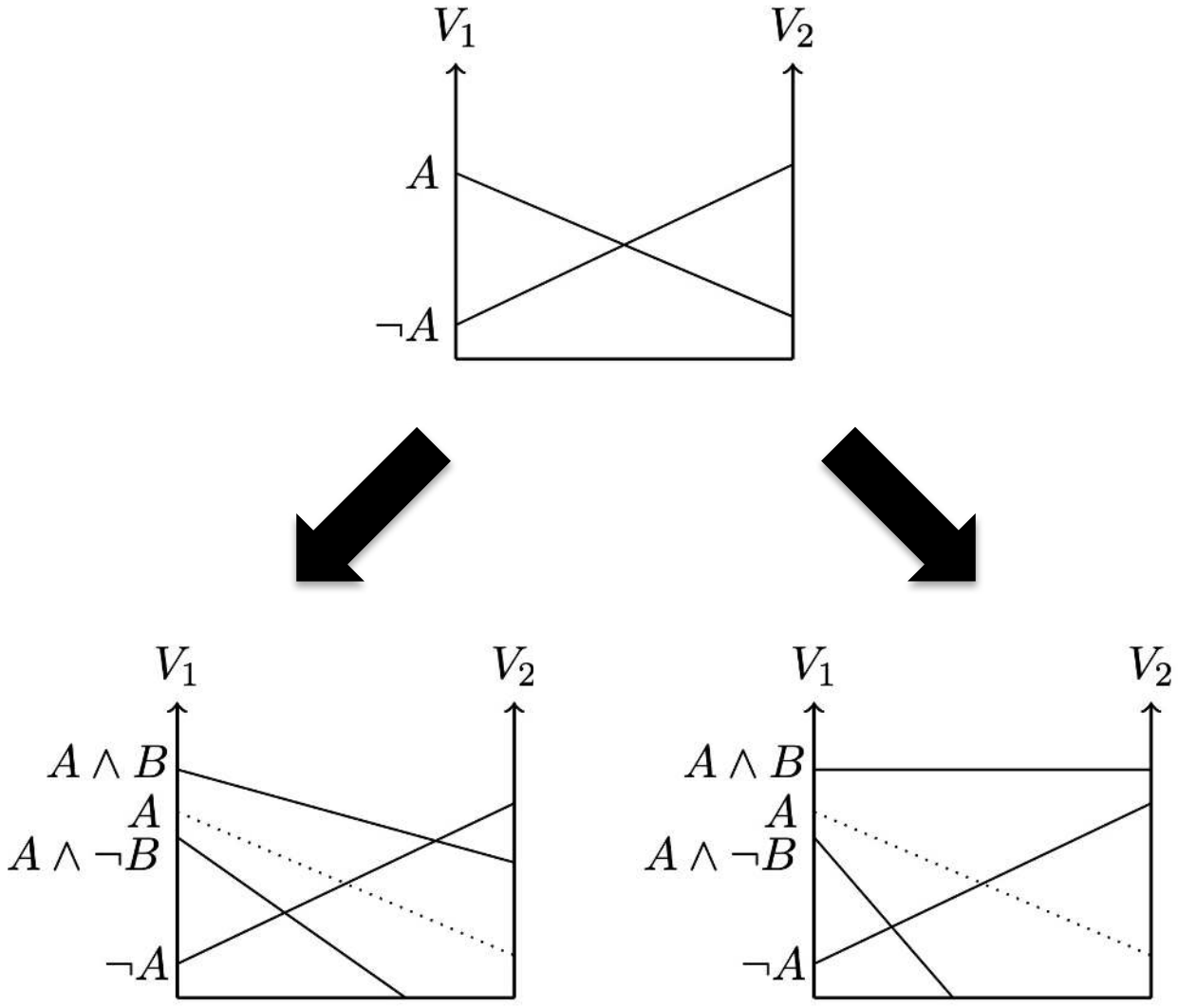
Refinement can help us with choice without commensuration by revealing **multi-value dominant options**.

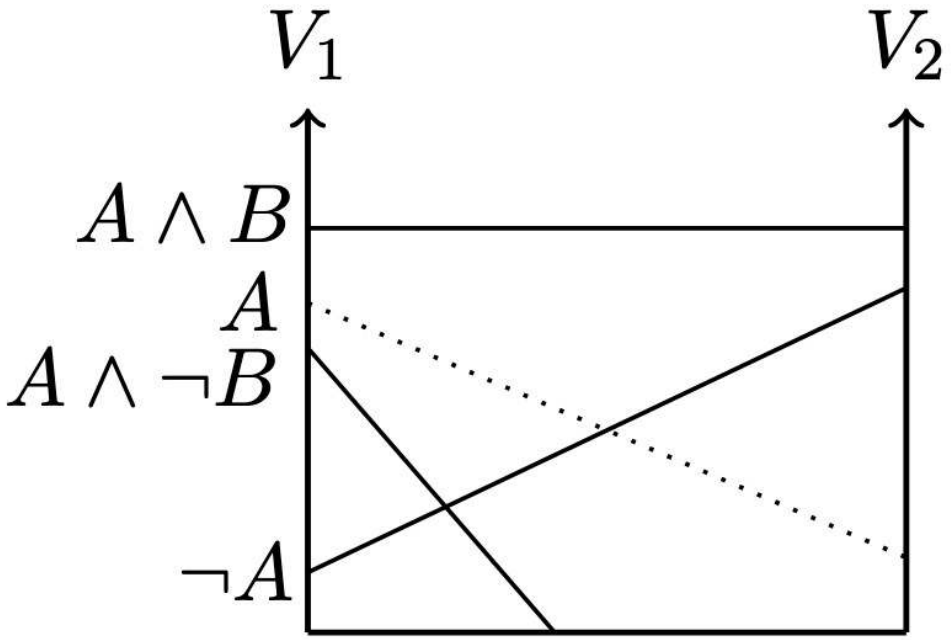

V_1
V_2
A
$\neg A$


V_1
V_2
A
$\neg A$




V_1
V_2
$A \wedge B$
A
$A \wedge \neg B$
$\neg A$


V_1
V_2
A
$\neg A$
V_1
V_2
$A \wedge B$
A
$A \wedge \neg B$
$\neg A$
V_1
V_2
$A \wedge B$
A
$A \wedge \neg B$
$\neg A$


V_1
V_2
$A \wedge B$
A
$A \wedge \neg B$
$\neg A$

Zero-Sum Games

Consider a simple 2×2 zero-sum strategic game.

	A^2	$\neg A^2$
A^1	$(v, -v)$	$(\alpha, -\alpha)$
$\neg A^1$	$(\beta, -\beta)$	$(\gamma, -\gamma)$

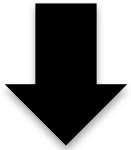
Base Game G_0

Refinement here corresponds to a player refinement one of her acts A to yield $A \wedge B_1$ and $A \wedge B_2$.

This yields a new 3×2 game with perturbed payoffs.

	A^2	$\neg A^2$
A^1	$(v, -v)$	$(\alpha, -\alpha)$
$\neg A^1$	$(\beta, -\beta)$	$(\gamma, -\gamma)$

Base Game G_0



	A^2	$\neg A^2$
$A^1 \wedge B_1$	$(v + \epsilon^1_{11}, -v + \epsilon^2_{11})$	$(\alpha + \epsilon^1_{21}, -\alpha + \epsilon^2_{21})$
$A^1 \wedge B_2$	$(v + \epsilon^1_{12}, -v + \epsilon^2_{12})$	$(\alpha + \epsilon^1_{22}, -\alpha + \epsilon^2_{22})$
$\neg A^1$	$(\beta, -\beta)$	$(\gamma, -\gamma)$

Refined Game G_ϵ

	A^2	$\neg A^2$
$A^1 \wedge B_1$	$(v + \epsilon^1_{11},\ -v + \epsilon^2_{11})$	$(\alpha + \epsilon^1_{21},\ -\alpha + \epsilon^2_{21})$
$A^1 \wedge B_2$	$(v + \epsilon^1_{12},\ -v + \epsilon^2_{12})$	$(\alpha + \epsilon^1_{22},\ -\alpha + \epsilon^2_{22})$
$\neg A^1$	$(\beta,\ -\beta)$	$(\gamma,\ -\gamma)$

Refined Game G_ϵ

The refinement reflection principle constraints the perturbations to satisfy $\mathbb{E}_{\mu}\left[\epsilon_{jk}^{i}\right] = 0$

That is, neither player expects that the sum of the payoffs of the refined subgame will be better or worse.

We’re interested in what happens to the aggregate welfare of the agents’ in the refined game.

Denote the Nash equilibria of the game G_k by $NE(G_k)$ with $k \in \{0,1\}$ for the pre- and post-refinement games,

Then the welfare-optimal equilibrium for a game is:

$$W_k^* = \max_{\sigma \in NE(G_k)} [U_1{}^k(\sigma) + U_2{}^k(\sigma)]$$

Since G_0 is zero-sum $W_0^* = 0$.

Theorem (Zero-Sum Escape). In the 2×2 zero-sum strategic game consider before, if both players players μ_A regarding refinement satisfy RRP and belief that payoff perturbations are not anti-correlated, then:

$$\mathbb{E}[W_1^*] > W_0^*$$

Intuition: refinement generates four new pure-strategy and, in expectation, some of these are positive sum.

	A^2	$\neg A^2$
$A^1 \wedge B_1$	$(v + \epsilon^1_{11},\ -v + \epsilon^2_{11})$	$(\alpha + \epsilon^1_{21},\ -\alpha + \epsilon^2_{21})$
$A^1 \wedge B_2$	$(v + \epsilon^1_{12},\ -v + \epsilon^2_{12})$	$(\alpha + \epsilon^1_{22},\ -\alpha + \epsilon^2_{22})$
$\neg A^1$	$(\beta,\ -\beta)$	$(\gamma,\ -\gamma)$

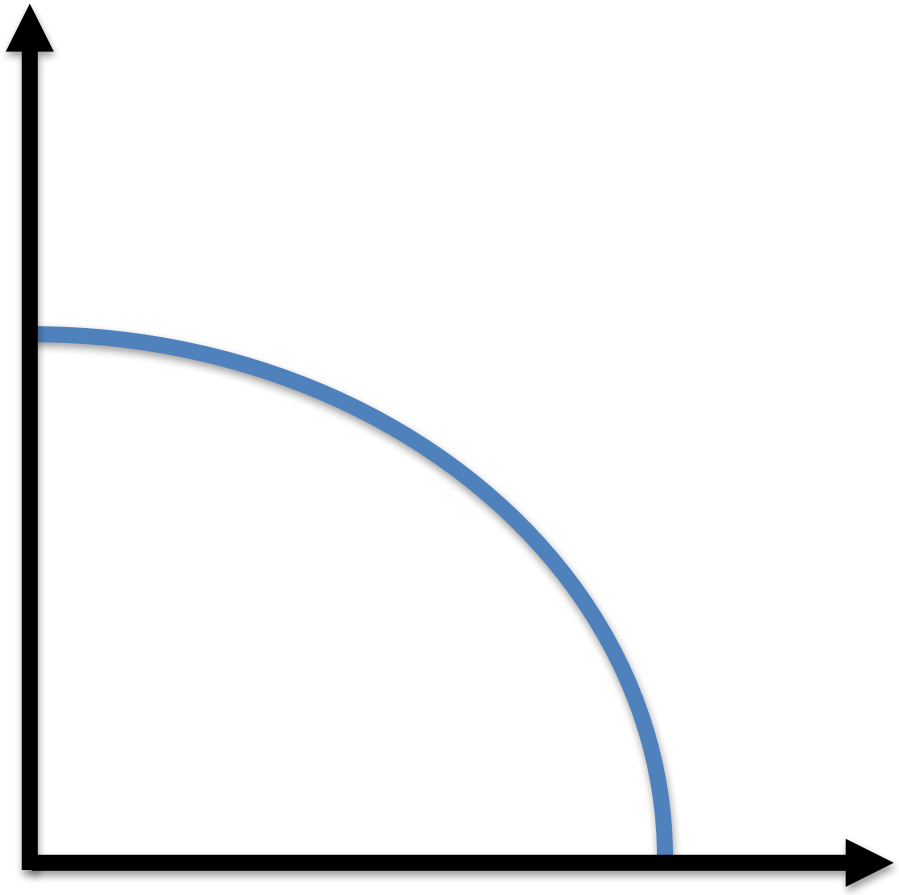
Refined Game G_ϵ

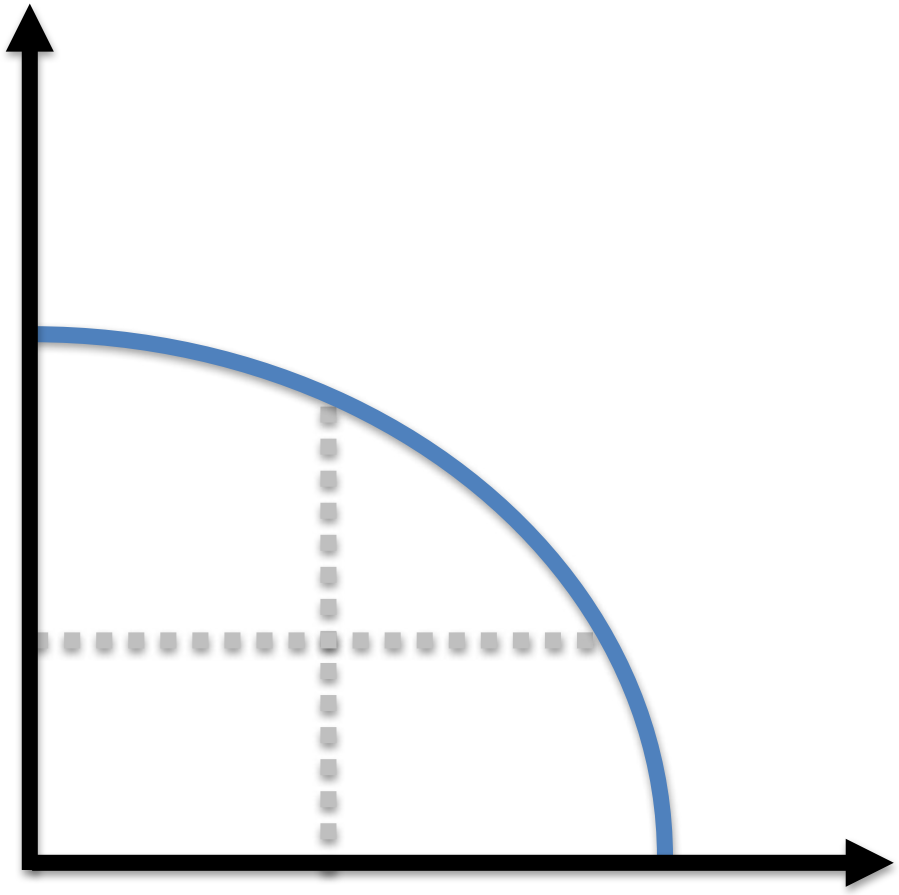
...and these new favorable profiles in turn will tend to composes elements of some mixed-strategy Nash equilibria of the game.

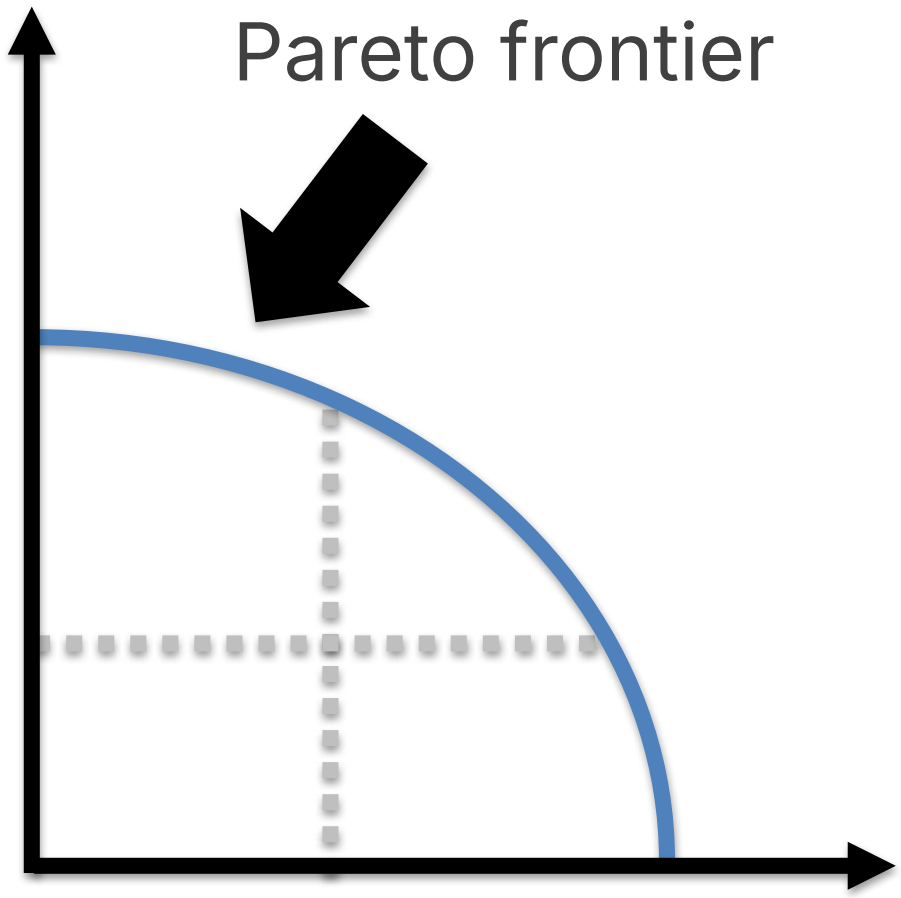
This show that apparent zero-sum conflicts often contain hidden positive-sum resolutions.

The finer-grained representation reveals opportunities for mutual benefit invisible at the coarse level.

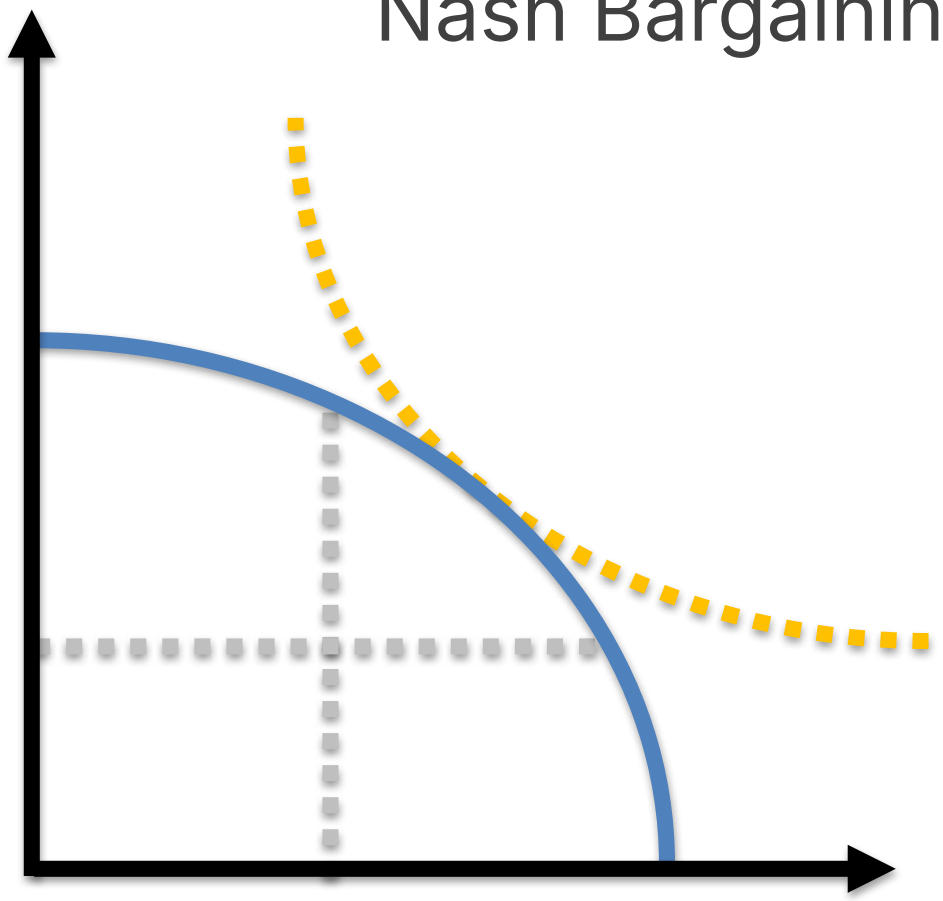
Nash Bargaining Games



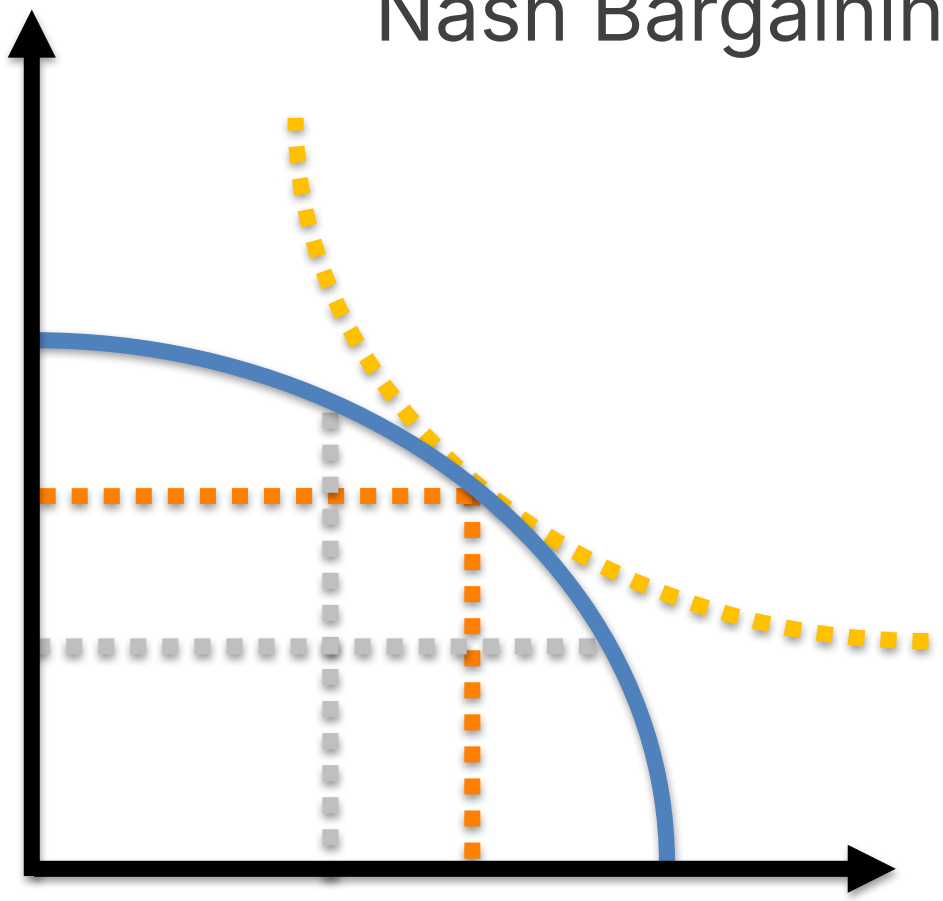


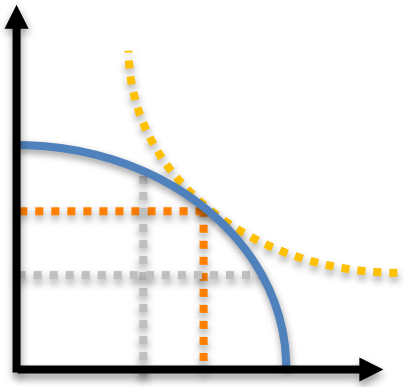

Pareto frontier

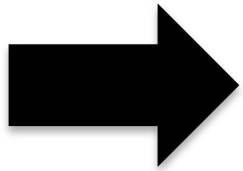
Nash Bargaining Solution



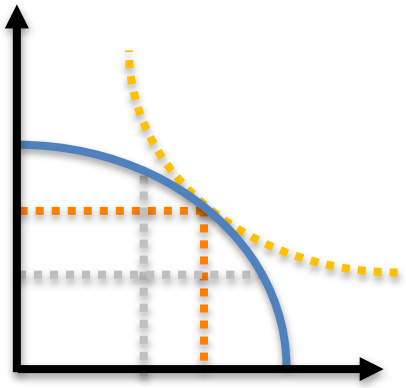
Nash Bargaining Solution



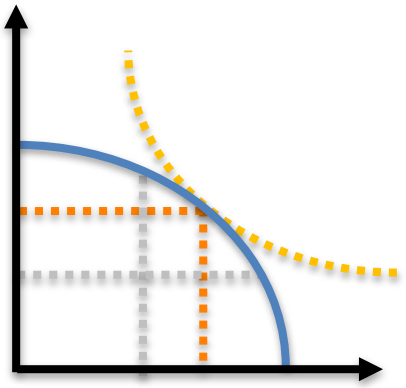


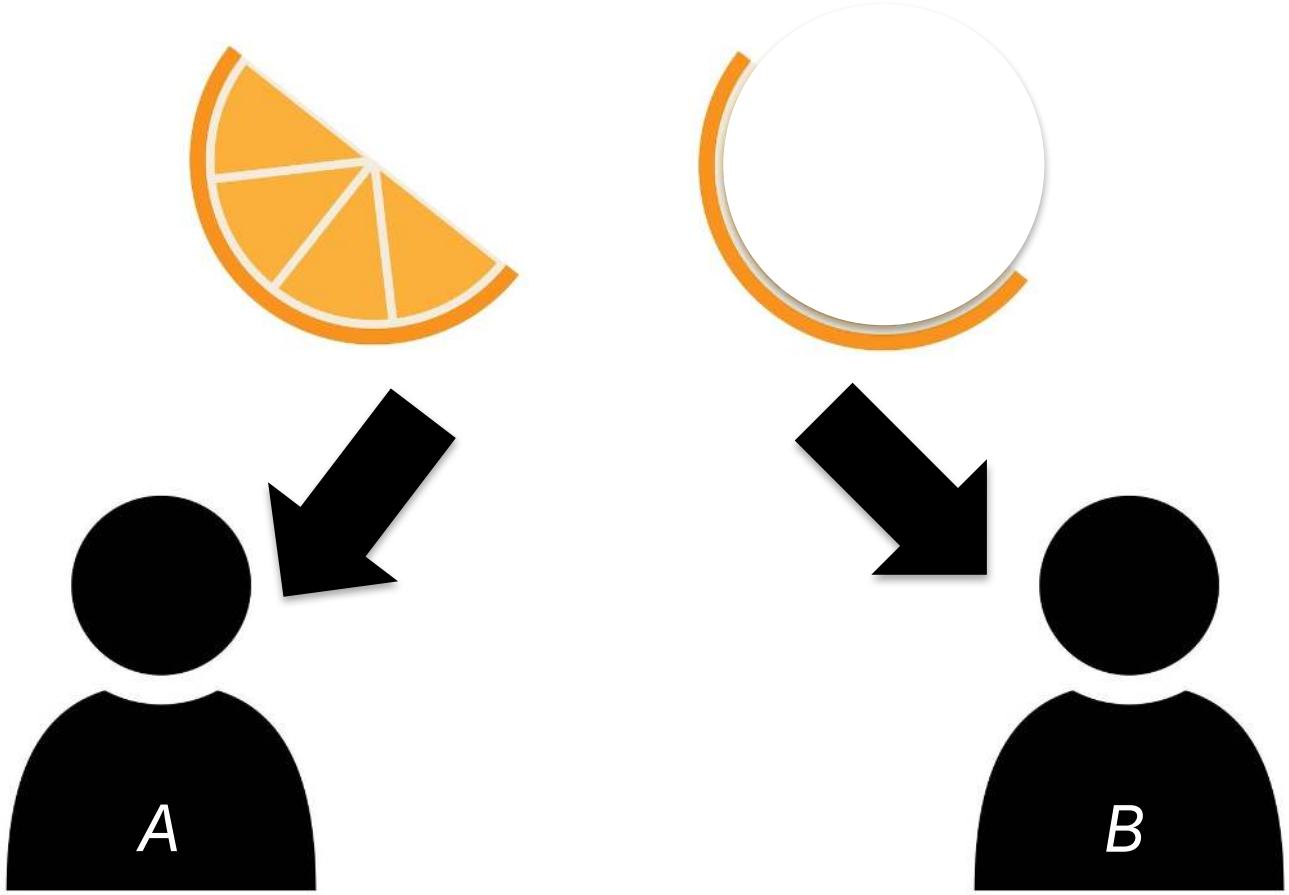


λ_i



$1 - \lambda_i$




A
B

Theorem (Value-Refinement in Nash Bargaining). In a two-player, symmetric Nash bargaining game, if players credences μ regarding refinement satisfy RRP and refinement uncertainty, then:

$$\mathbb{E}[W_1^*] > W_0^*$$

That is, the payoffs Nash payoffs post-refinement strictly pareto-dominate pre-refinement payoffs.