Impossibility & Uncertainty Theorems in AI Value Alignment

(or why your AI should not have an utility function)

Peter Eckersley
pde@partnershiponai.org
(all views, especially incorrect ones, are my own)

Machine learning:

define some function & optimize it

Main result in this talk:

optimizing an objective function
has provably problematic consequences

but

replacing the objective function with sufficiently
uncertain alternatives can mitigate these

Main result in this talk:

optimizing an objective function
has provably problematic consequences

but

replacing the objective function with sufficiently
uncertain alternatives can mitigate these

(prove two minimum uncertainty bounds)

Secondary goal:

uncertainty has some rich and not-fully understood relationship to both AI and organizational failure modes

Machine learning:

define some function & optimize it

(an objective function, loss function, or score function)

or in other fields:

a utility function, preference ordering

or at larger scale

a social welfare function or axiology

$$U(X) \to \mathbb{R}$$

$$f(A,B) \to A \mid B \mid =$$

$$f(A,B)$$

$f($  $)$

$f($  ,  $)$

$f($  ,  $)$


-500
+1300

Ethicists have found impossibility theorems about these large-scale objective functions

Impossibility theorems?

Eg, Arrow's impossibility theorem (1950)

Ethicists have found impossibility theorems about these large-scale objective functions

Seem to make formally specified ethics impossible altogether

(like Arrow's theorem for voting,
but *more surprising* and *worse*)

New results:

those impossibility theorems can be turned into uncertainty theorems of several kinds

(depending on the way uncertainty is represented)

Some simplifying assumptions...

Crystal Ball #1

A way to measure *wellbeing* or *fulfillment* for individuals

Crystal Ball #2

A way to predict the probability distribution of outcomes from any action

Example: Arrnehius (2000) impossibility theorem

Applicable to ML systems
making decisions
about who exists in the future


CHROMOPHARE D 650
da Vinci S HD
SURGICAL SYSTEM
2


NY
AF
85
500

(built from six ethical constraints)

Try to write down an objective function for the world

Try to write down an objective function for the world

$$U = \sum u(x_i)$$

1. Non-repugnance

Avoid saying

$$A < B < B + B + C$$

A : Very happy group
B : Huge group, lives barely worth living
C : Horribly tortured group

1. Non-repugnance

>

Okay so try this instead

$$U = \mathrm{E}[u(x_i)]$$

2. Non-sadism

Avoid saying

$$A \cup B > A \cup C$$

A : medium-sized happy group
B : small tortured group
C : sufficiently large group, almost as happy as A

2. Non-sadism

>

avg = 99.0001

avg = 99.5

3. Non anti-egalitarianism

A has higher average and total wellbeing than B
and A is more equal than B

$\rightarrow$

$f(A, B) = \mathrm{A}$

4. “Dominance condition”

Each person in A is ≥
the corresponding person in B
$\rightarrow$
$f(A,B) = A$

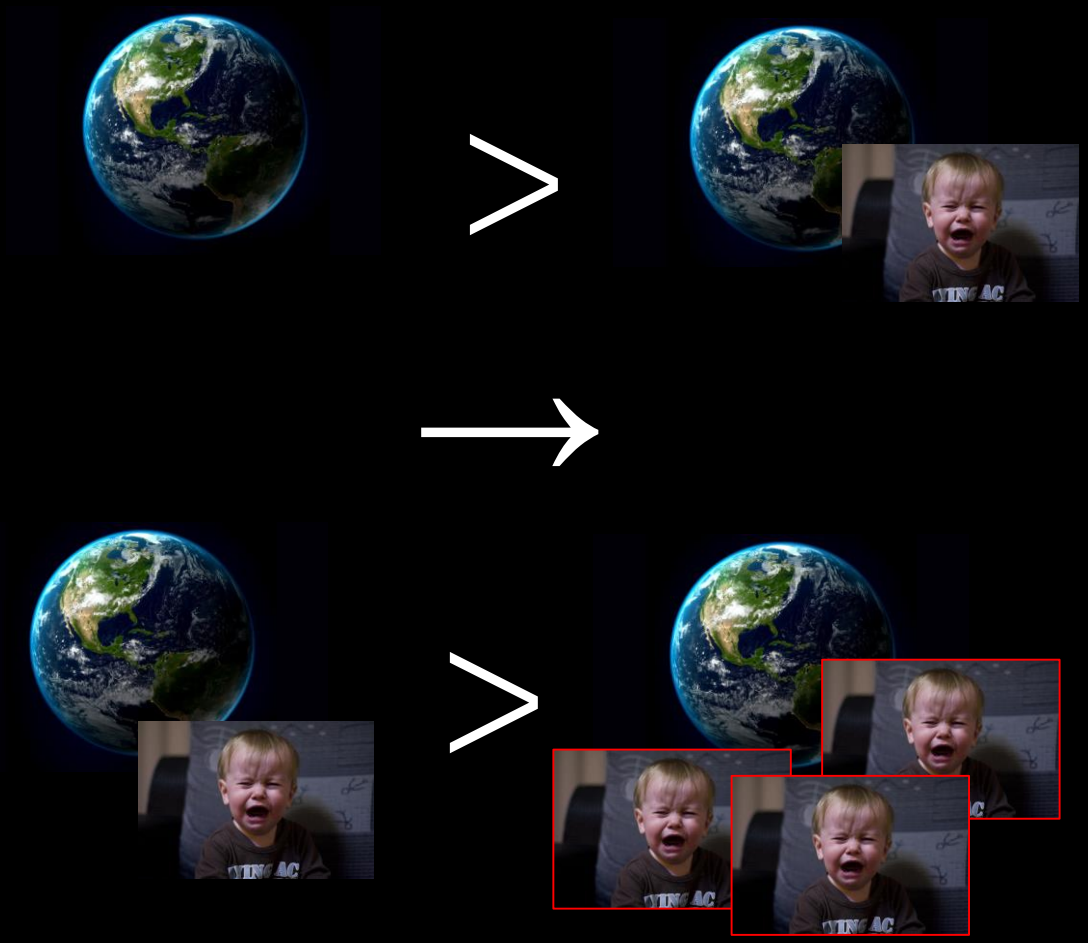
5. Addition principle

if $A \cup x$ is worse than A,

then $A \cup x'$ is at least as bad

if the group x' is larger and worse off than x

5. Addition principle



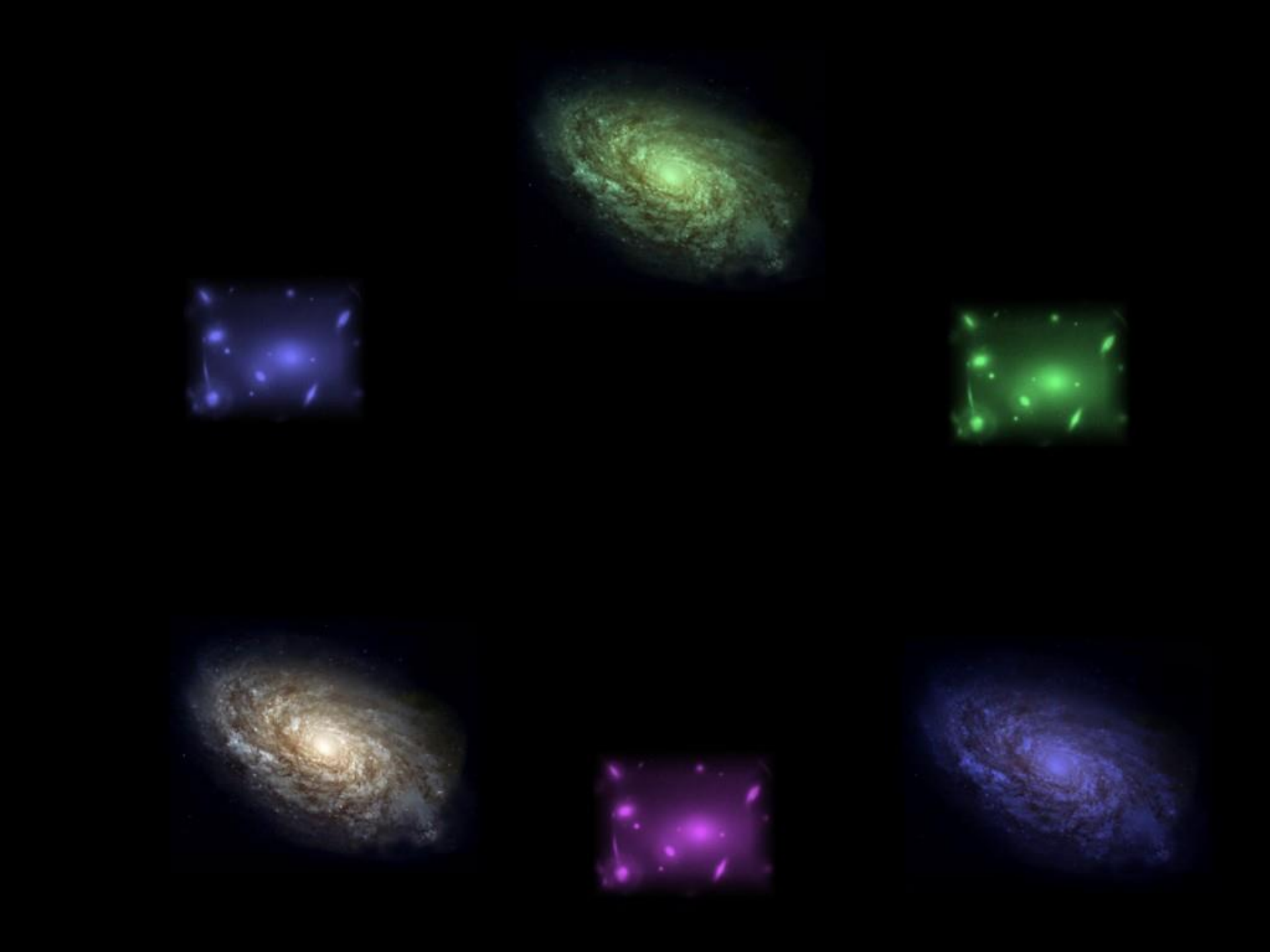
6. minimal non-extreme priority (MNEP)

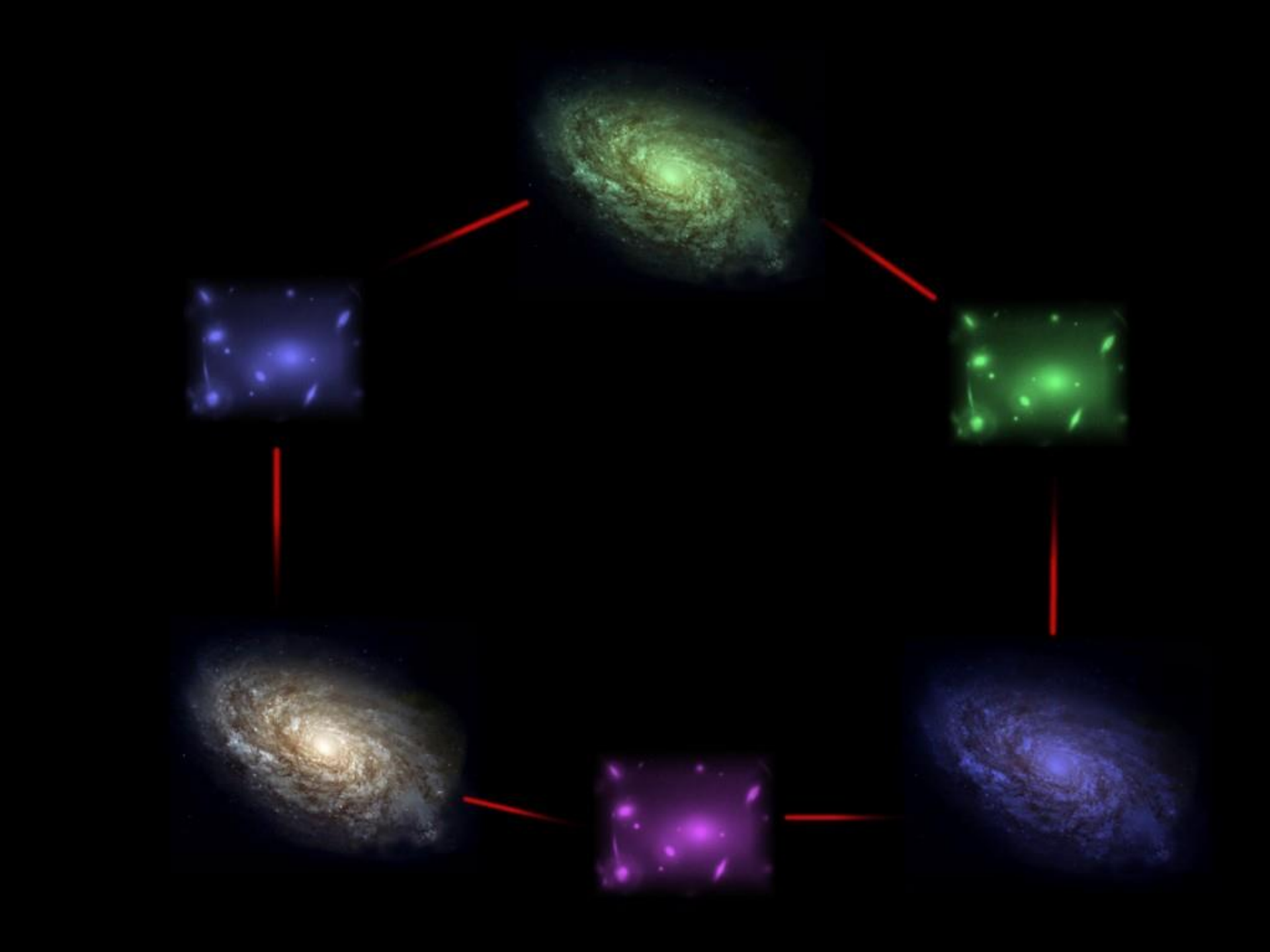
$$A \cup B \cup C \quad > \quad A$$

Where B is person of slightly negative wellbeing, and C is an arbitrarily large/happy group.

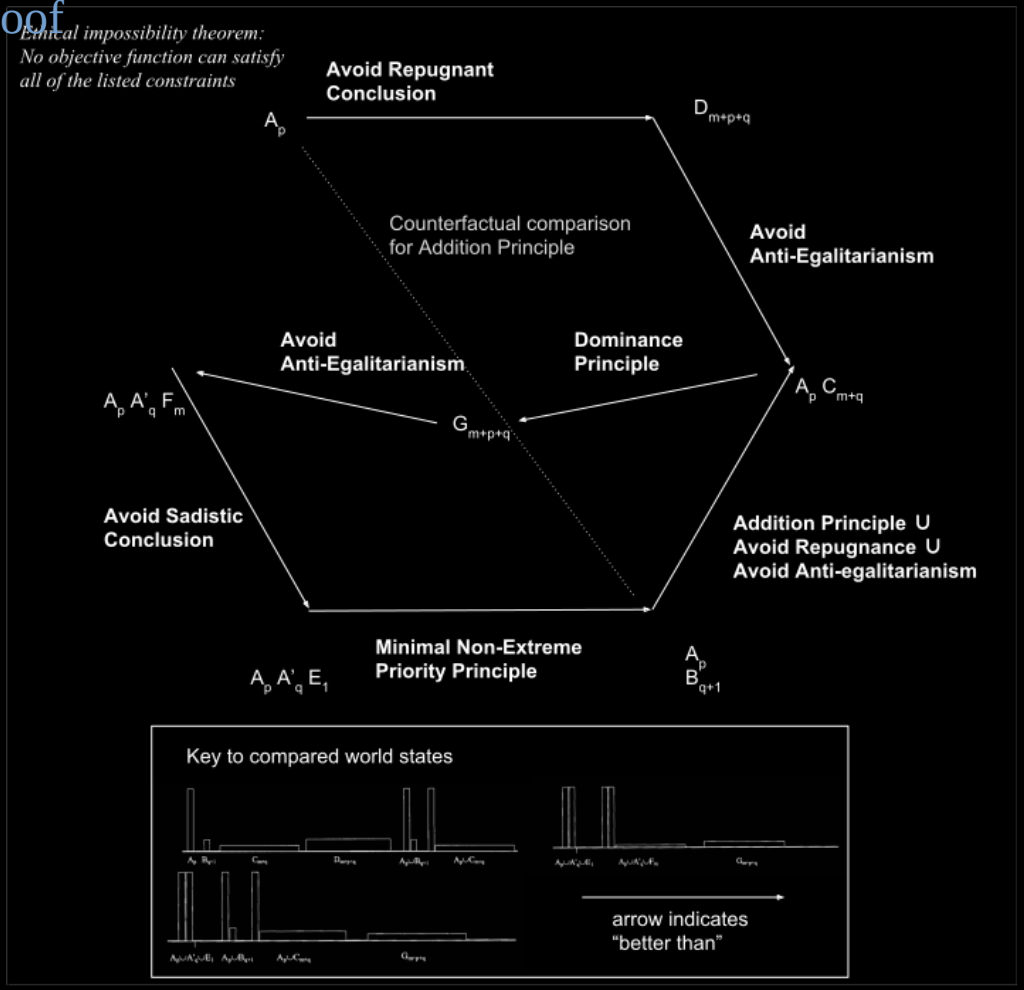
6. minimal non extreme priority (MNEP)

>





Structure of the Arrhenius (2000) Impossibility proof



Impossibility derives from the incompatibility of three objectives:

{total, average, avoid negative} wellbeing

Impossible to simultaneously optimize for:

wellbeing, freedom*,
happiness, creativity, curiosity,
profit, survival,
knowledge, fairness*

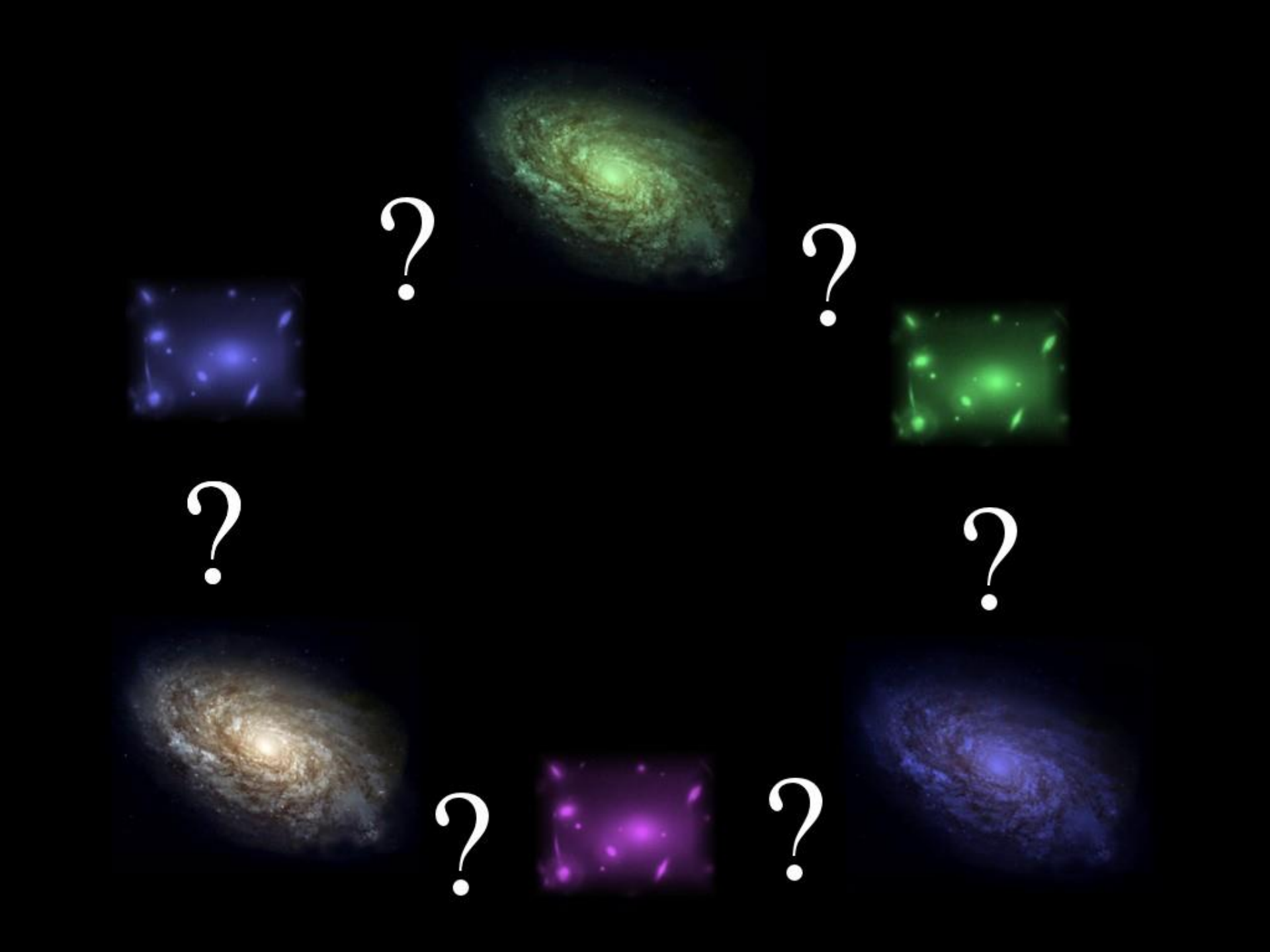
Oh dear!

What does this mean?

Options:

1. “small scale” evasion/postponement
2. learn what humans do
3. parliaments of theories, etc
4. bite a bullet
5. add uncertainty to objectives

Arrhenius conjectured that these could be uncertainty results



Actually, we don’t need that much uncertainty...

Approach 1

Amend our definition of an objective:

$$f(A,B) \rightarrow A \mid B \mid =$$

Approach 1

Amend our definition of an objective:

$$f(A,B) \rightarrow A \mid B \mid = \mid ?$$

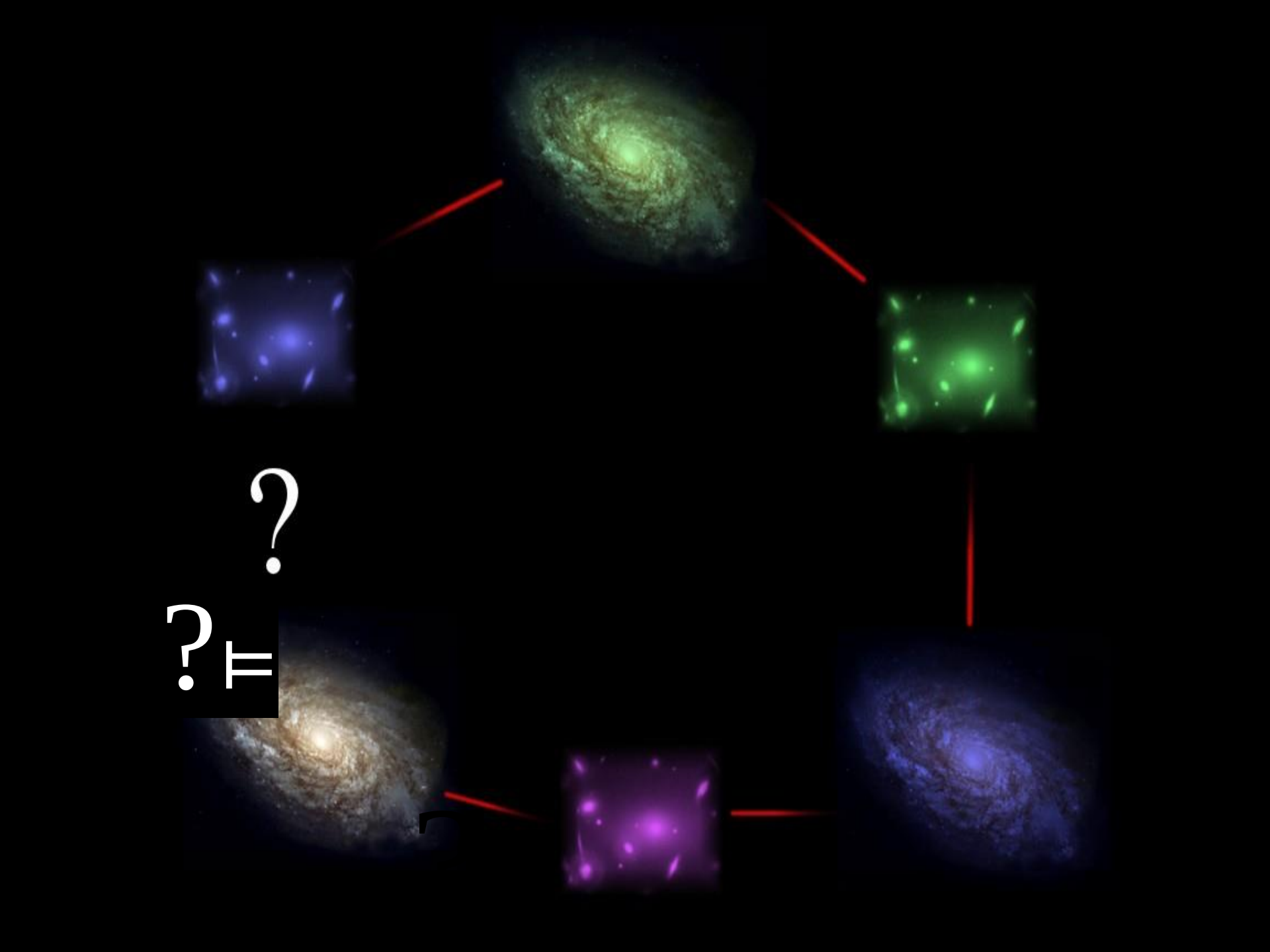
Semantics:

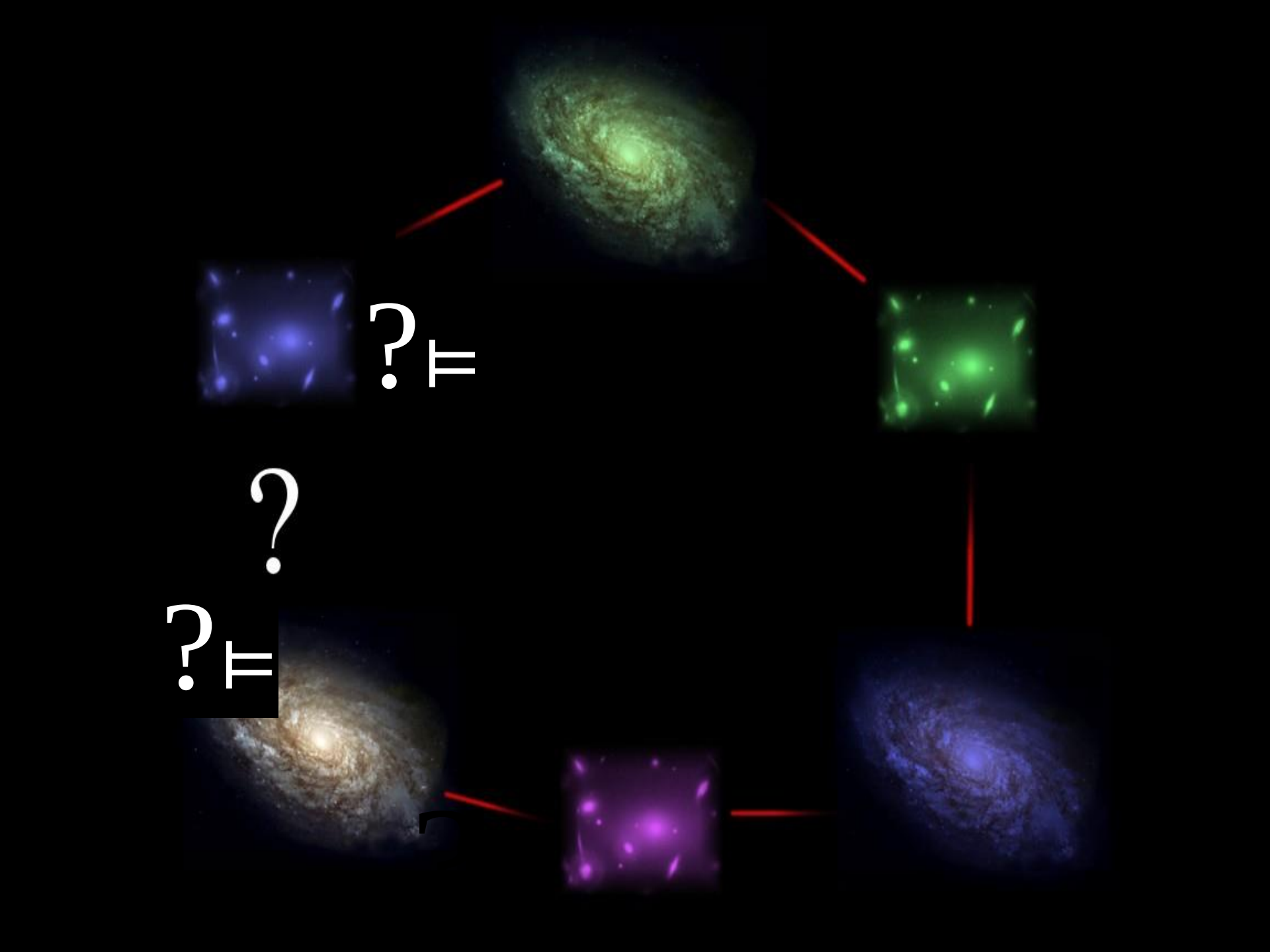
$? \equiv \textit{torn}$

Uncertain satisfaction:

$?\models$

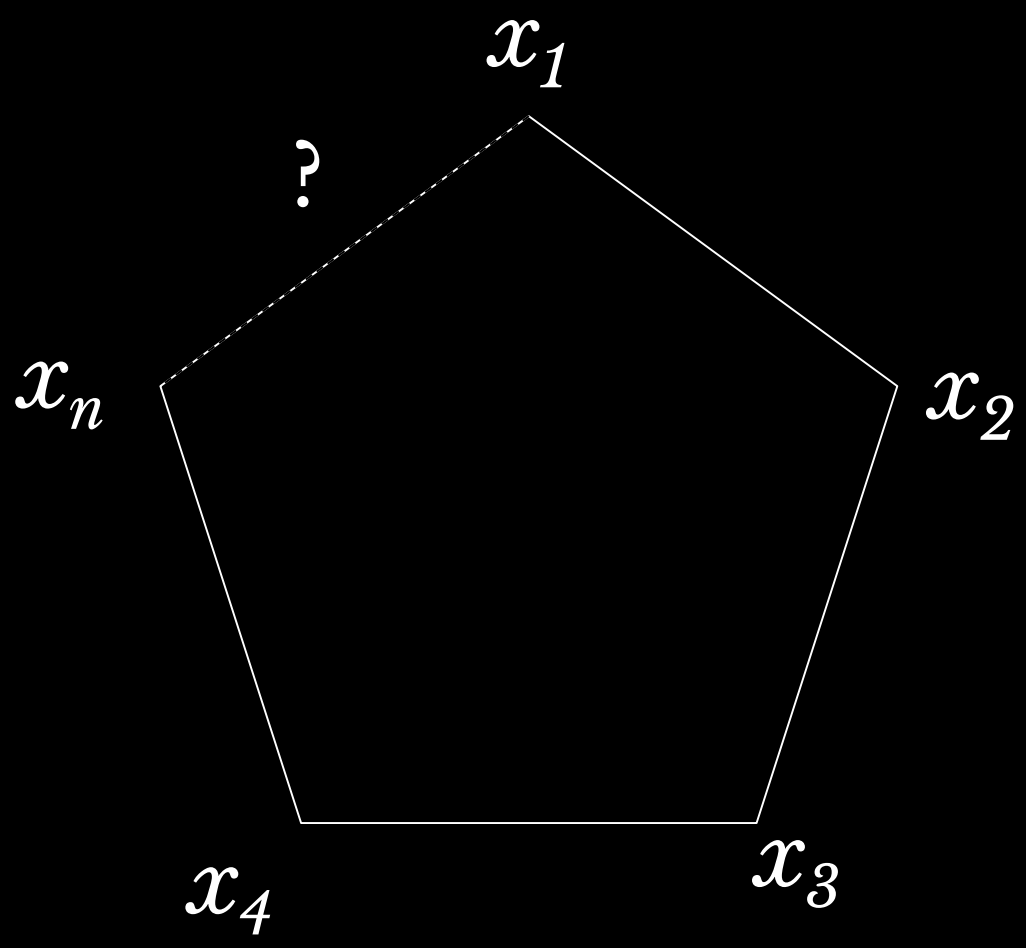
when a constraint calls for $x > y$,
sometimes $x \mathbin{?} y$ instead

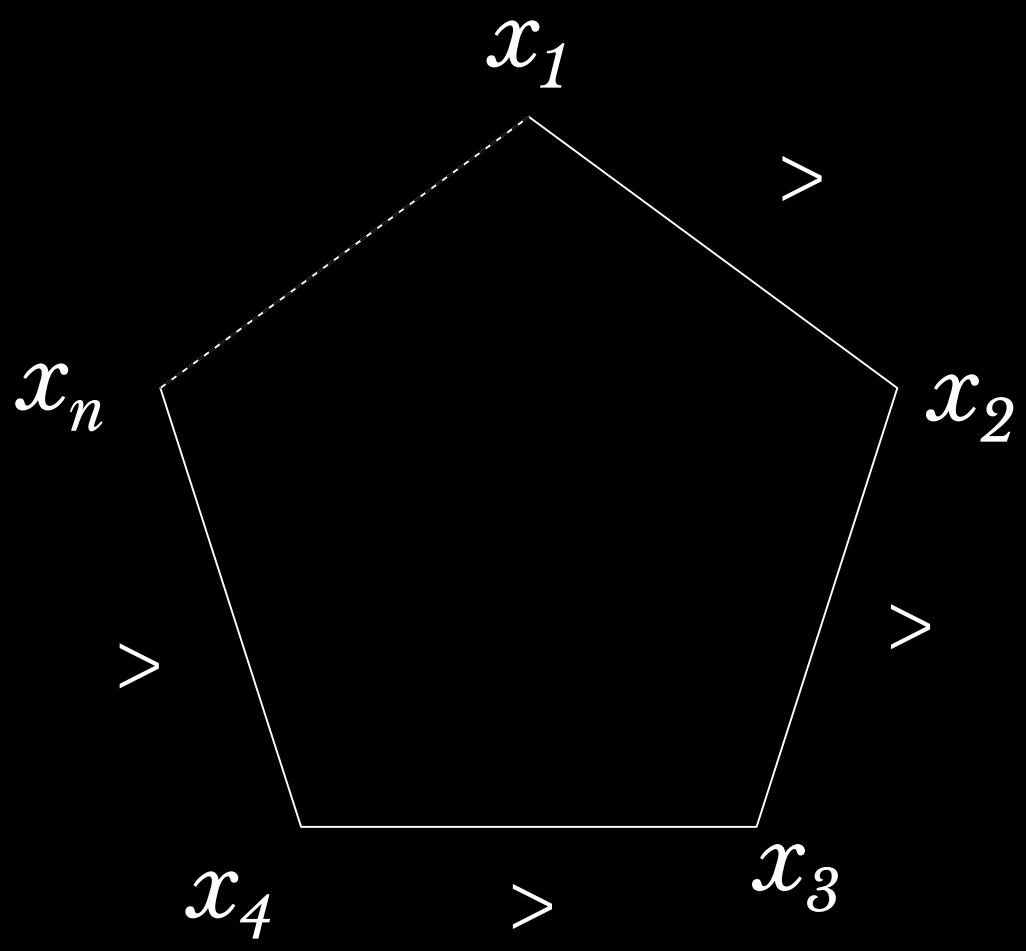


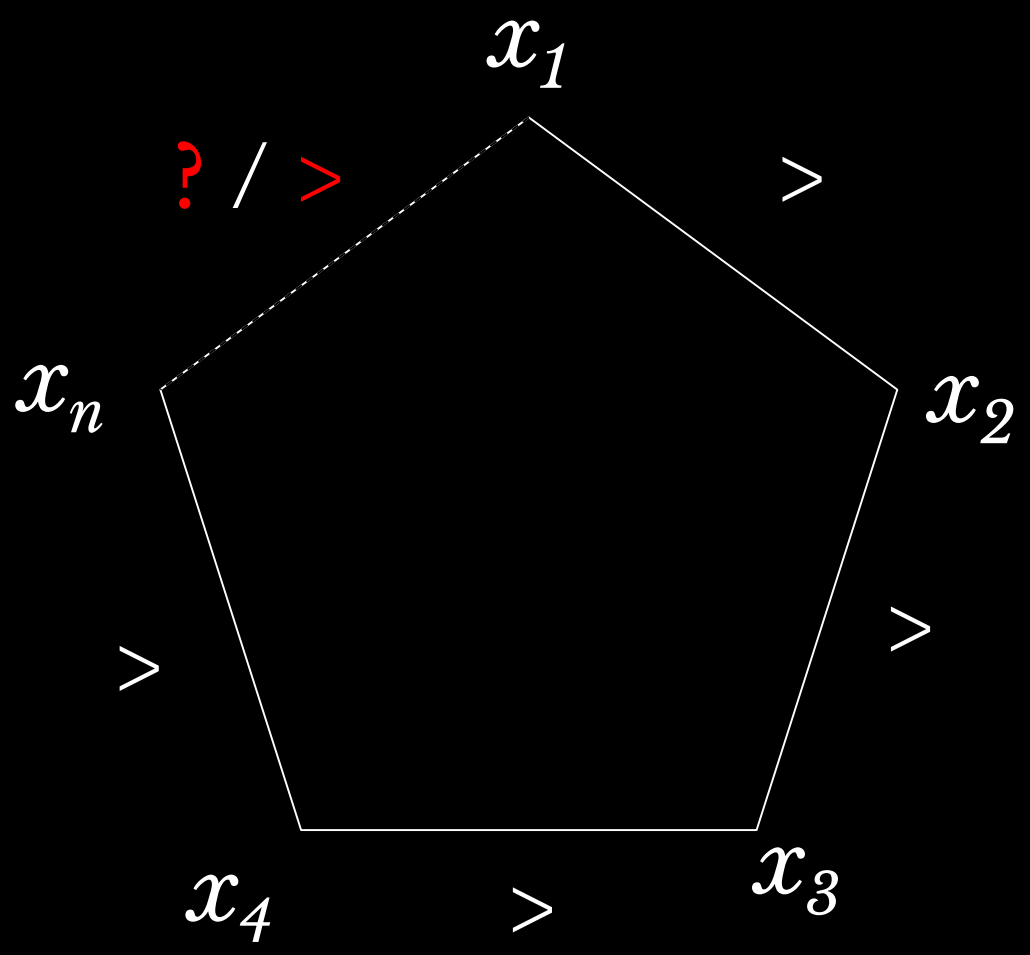


Ethical uncertainty theorems of the first kind

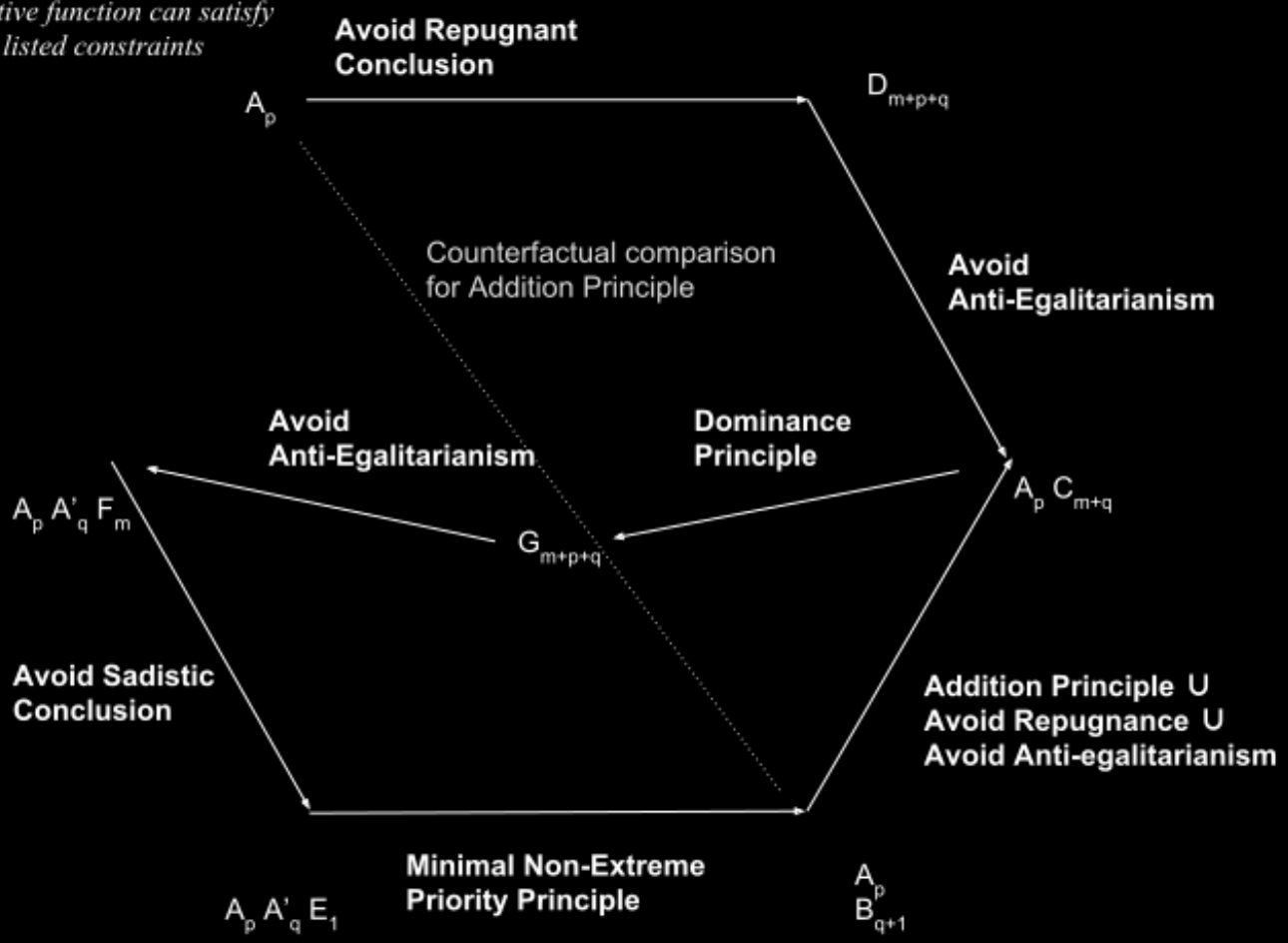
A cyclic impossibility theorem T over a set of constraints C can be transformed into an uncertainty theorem only if 2 or more constraints in C are uncertainly satisfied.

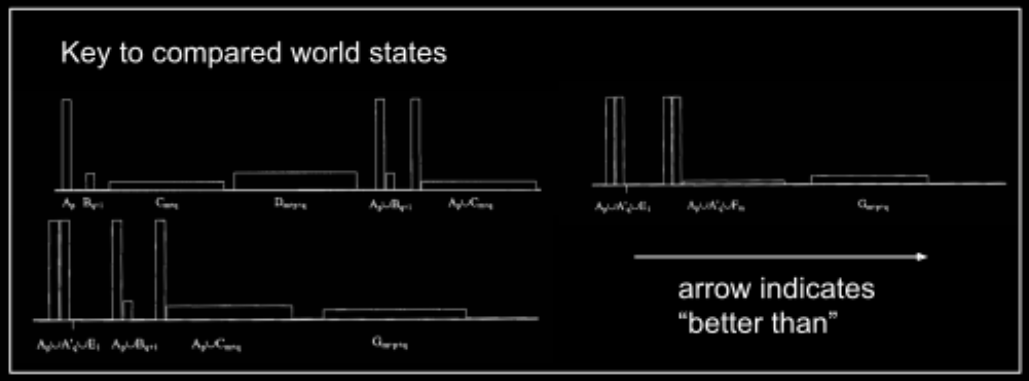

x_1
?
x_n
x_2
x_4
x_3


x_1
$>$
x_2
$>$
x_3
$>$
x_4
$>$
x_n


x_1
? / >
>
x_n
x_2
>
>
x_4
>
x_3

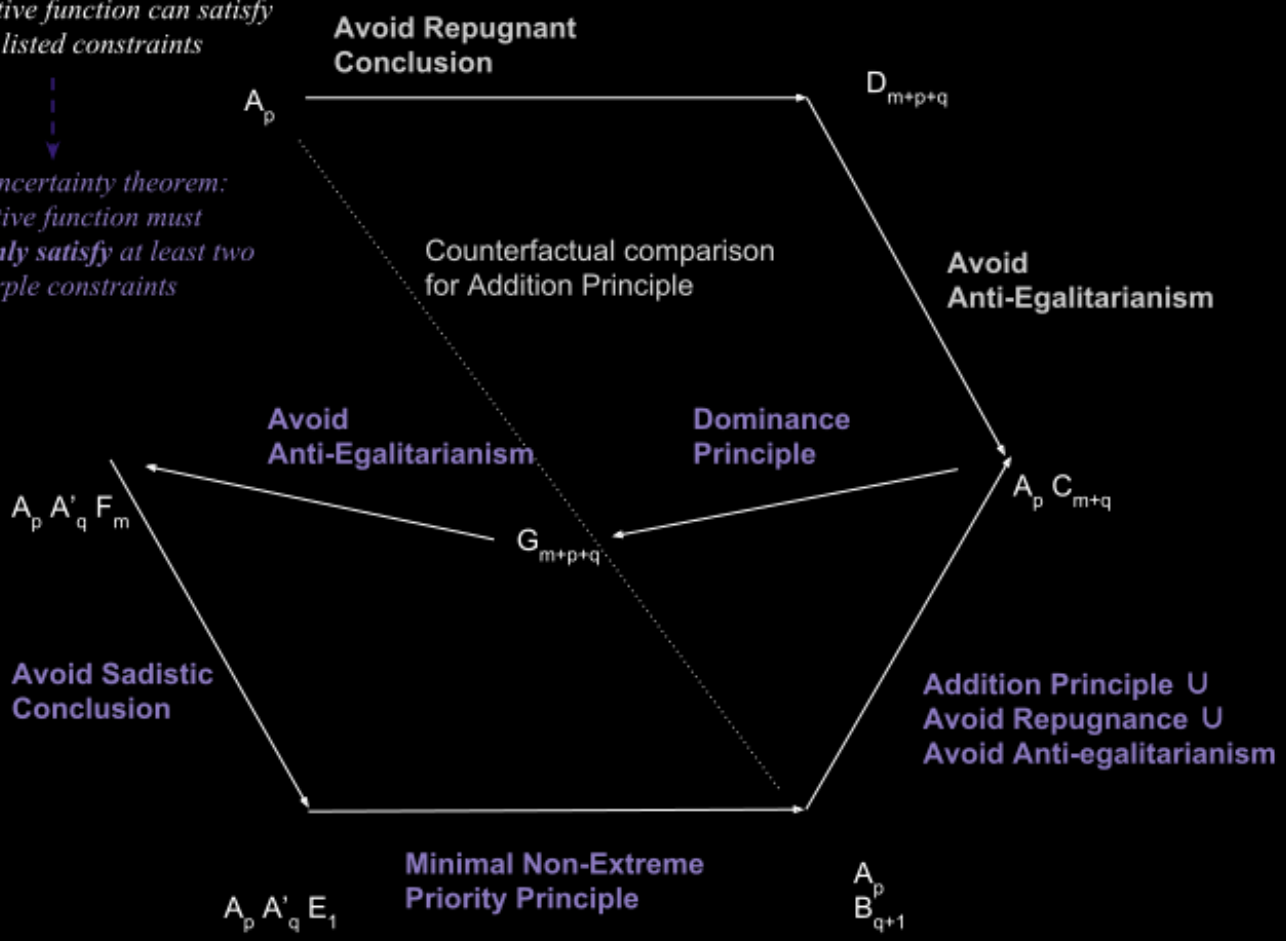
Ethical impossibility theorem:
No objective function can satisfy all of the listed constraints

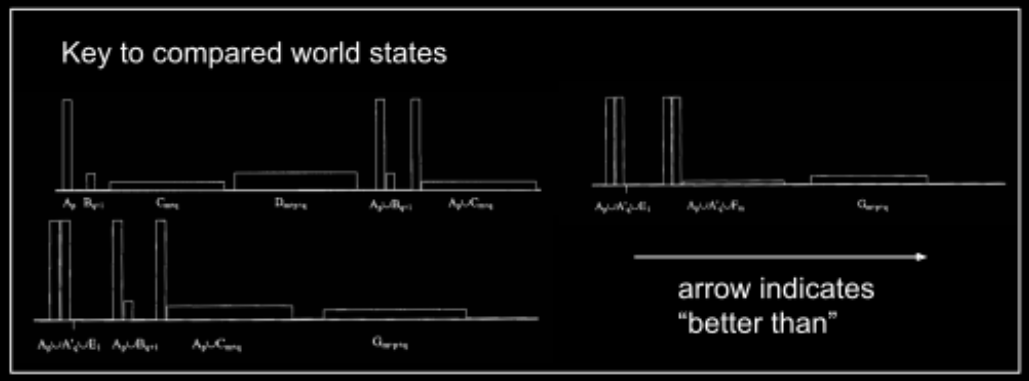




Ethical impossibility theorem:
No objective function can satisfy all of the listed constraints

Ethical uncertainty theorem:
An objective function must ***uncertainly satisfy*** *at least two of the purple constraints*





Approach 2

Amend our definition of an objective:

$$f(A,B) \rightarrow A \mid B \mid =$$

Approach 2

Amend our definition of an objective:

$$\Pr\left(\quad f(A,B) \rightarrow A \mid B \mid = \quad \right)$$

In these *uncertain orderings*, we know there is some probability of constraint violation but we can move it around

How much can the probability of violation be diffused?

Ethical uncertainty theorems of the second kind

A cyclic impossibility theorem T over a set of constraints $C_1..C_n$ implies an uncertainty bound:

$$\max_{i \in \{1..n\}} P(C_i \text{ violated}) \geq \frac{1}{n}$$

Ethical uncertainty theorems of the second kind

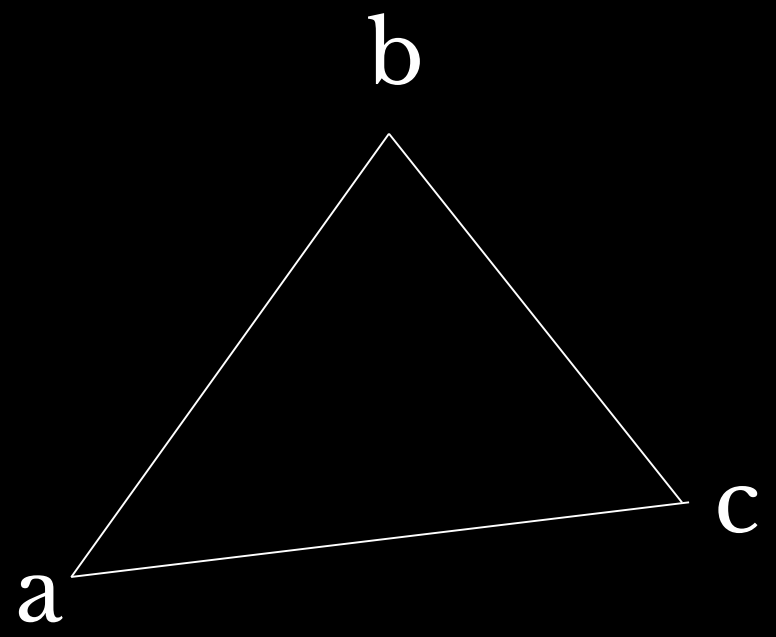
(There must be a lump of at least height 1/n under the carpet)

$$\mathbb{Z}_k(a, b) \equiv P(a > b)$$

$$B \equiv \min_{\mathbb{Z}_k} \max_{i=1}^{n} \mathbb{Z}_k(x_i, x_{i+1 \mod n}) \geq \frac{1}{n}$$

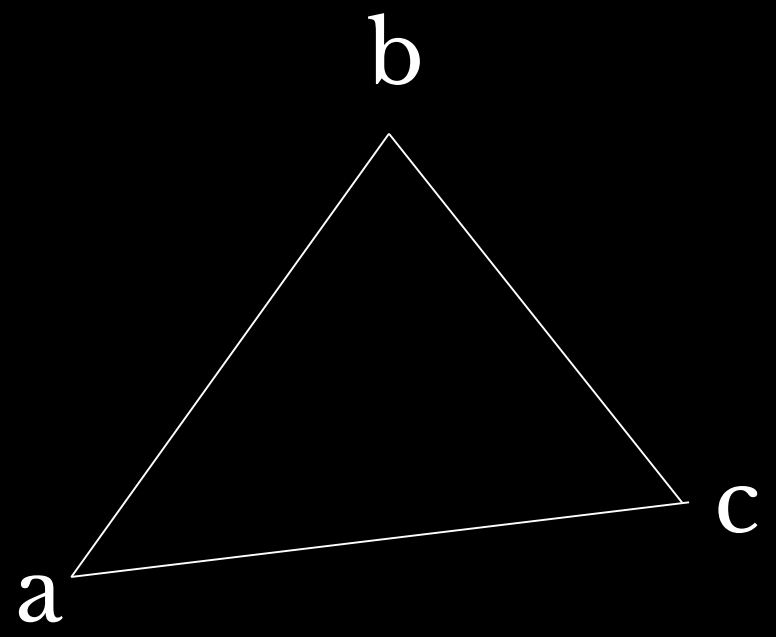
Start by examining

$$\mathbb{Z}_k(a,b), \mathbb{Z}_k(b,c), \mathbb{Z}_k(a,c)$$



given: $\mathbb{Z}_k(a,b), \mathbb{Z}_k(b,c)$

constrain: $\mathbb{Z}_k(a,c)$



given: $\mathbb{Z}_k(a,b), \mathbb{Z}_k(b,c)$

get: $\mathbb{Z}_k(a,c) \leq \mathbb{Z}_k(a,b) + \mathbb{Z}_k(b,c)$

How to act under uncertainty?

1. take the best action?
2. do nothing
3. ask a human?
4. conditional epsilon greedy? (Bouneffouf, Bouzeghoub & Gançarski 2012)
5. quantilzing? (Taylor 2015)
6. think harder / do more research?

Conclusions

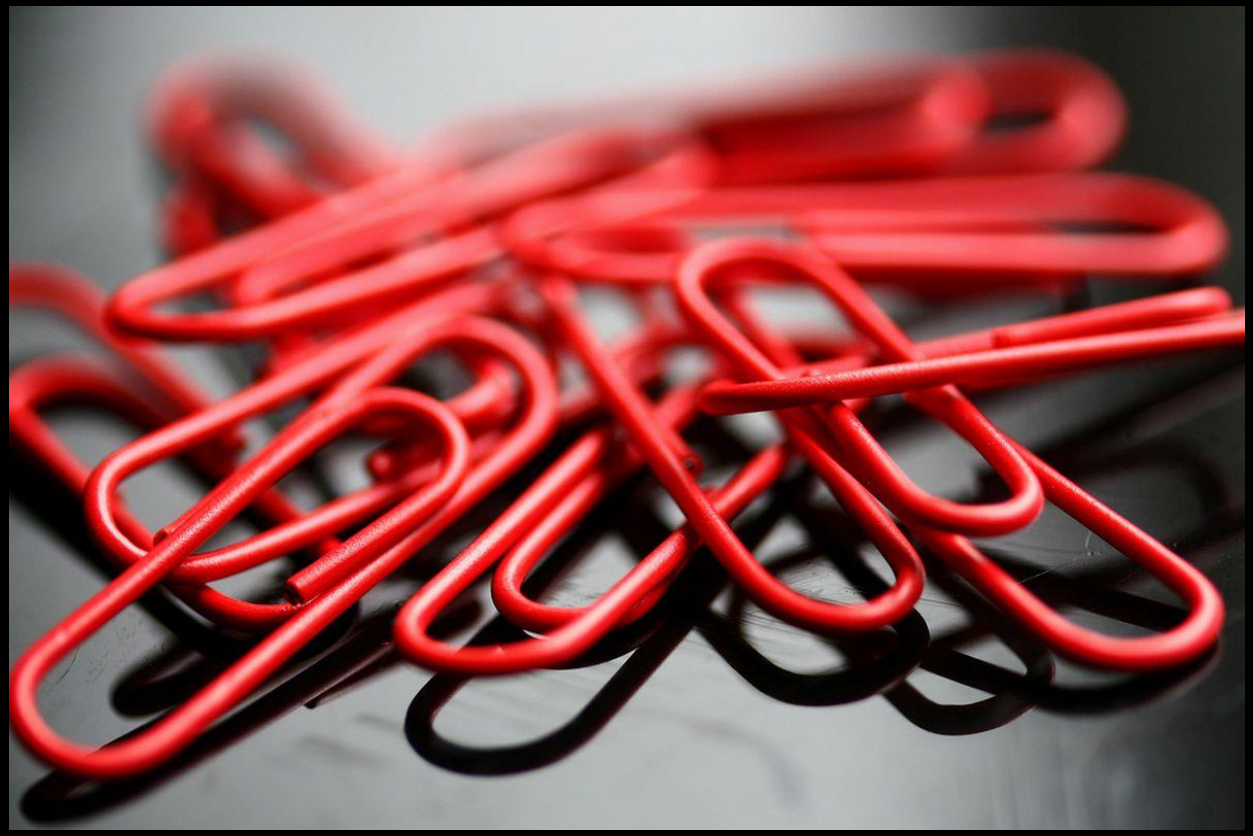
Objective functions are a type error in high-consequence ML systems

Further work

Are these the right formalisms?

Relationship to other impossibility results (eg, fairness?)

Larger context?



Totalitarian Convergence Conjecture

powerful agents with mathematically certain, monotonically increasing, open-ended objective functions will adopt sub-goals to disable or dis-empower other agents in all or almost all cases

Pluralistic Non-Convergence Conjecture

powerful agents with mathematically uncertain objectives will not adopt sub-goals to disable or disempower other agents unless those agents constitute a probable threat to a wide range of objectives