Falsifiability of theories of deliberated preferences

Olivier Cailloux [1]

[1]LAMSADE, Université Paris-Dauphine, PSL

4th December, 2023

Dauphine UNIVERSITÉ PARIS | PSL

LAMSADE UMR CNRS 7243 laboratoire d'analyse et modélisation de systèmes pour l'aide à la décision

Outline

Outline

1. Deliberated judgments and preferences
2. Theories of deliberated preference
3. Properties and existence of theories
4. Discussion

Deliberated judgment

- Individual i wonders about some issue
- Possible judgments J (e.g. yes/no, beautiful / ugly / neutral, ...)
- Shallow judgment: the one without arguments
- Deliberated judgment: the one that is stable facing counter-arguments
- Represents the judgment after having considered all arguments from a given set of arguments

Deliberated preference

- Individual *i* wonders about choosing some option among two possibilities
- Possible preferences $\{\varphi, \neg\varphi, 0\}$ meaning “pick first option”, “pick second option”, “no preference”
- Examples: beer VS milkshake, vegan meal VS meat, teaching using flipped classroom VS not, . . .
- Shallow preference: the one without arguments
- Deliberated preference: the one that is stable facing counter-arguments
- Represents the preference after having considered all arguments from a given set of arguments

Formal context

- Options $P = \{\varphi, \neg\varphi, 0\}$
- Individuals I
- Arguments $\mathscr{A} = \{a_1, \dots\}$
- Attitude $\rightsquigarrow$: the reactions of individuals to arguments (unknown but partially observable)

Example: flipped classroom

- Options $P = \{\varphi =$ "flipped classroom" $= flp, \neg\varphi =$ "classical approach", $0 =$ "no preference"$\}$
- Individuals I: the teachers in this room
- Arguments $\mathscr{A}$: a set of fifty arguments about or against flipped classrooms (studies, personal experience, . . .)
- Attitude $\rightsquigarrow$: however the teachers react to the arguments

Sequence of arguments

- $\alpha \in \mathscr{A}^{<\mathbb{N}}$: a finite sequence of arguments
- $\alpha \rightsquigarrow_i \varphi$: individual i after seeing α (in order) opts for φ (also $\rightsquigarrow_i(\alpha) = \varphi$)
- Attitude $\rightsquigarrow \in P^{\mathscr{A}^{<\mathbb{N}}I} = \{\rightsquigarrow_i \mid i \in I\}$

Example: attitude

- $\emptyset \rightsquigarrow_{\text{Alexis}} \mathit{flp}$: Alexis opts for *flp* without arguments
- $(a_1) \rightsquigarrow_{\text{Alexis}} \neg \mathit{flp}$: Alexis rejects *flp* if given a_1
- $(a_1, a_2) \rightsquigarrow_{\text{Alexis}} \mathit{flp}$: Alexis opts for *flp* if given a_1 then a_2
- $(a_1, a_2) \rightsquigarrow_{\text{Olivier}} \mathit{flp}$, $(a_2, a_1) \rightsquigarrow_{\text{Olivier}} \neg \mathit{flp}$: Olivier opts for *flp* if given a_1 then a_2 but not the other way around

$\rightsquigarrow$ encodes the reactions of all individuals to every possible sequence of arguments

Decisive argument

Decisive argument

a is *decisive* for i in favor of φ iff it convinces i whenever it appears within the last two arguments:

$$a \hookrightarrow_i \varphi \iff \forall \alpha \mid a \in \alpha_{[\![\#\alpha-1,\#\alpha]\!]} : \alpha \rightsquigarrow_i \varphi$$

Uniqueness

If a is decisive for i in favor of φ, there is no decisive argument for i in favor of any $p \neq \varphi$

Example: decisive argument

- Is a_1 decisive for Olivier?
- Is a_2 decisive for Alexis?

Decisive argument

Decisive argument

a is *decisive* for i in favor of φ iff it convinces i whenever it appears within the last two arguments:

$$a \hookrightarrow_i \varphi \iff \forall \alpha \mid a \in \alpha_{[\![\#\alpha-1,\#\alpha]\!]} : \alpha \rightsquigarrow_i \varphi$$

Uniqueness

If a is decisive for i in favor of φ, there is no decisive argument for i in favor of any $p \neq \varphi$

Example: decisive argument

- Is a_1 decisive for Olivier? No (not in favor of 0 or $\neg flp$ as $(a_1, a_2) \rightsquigarrow_i flp$ and not in favor of flp as $(a_2, a_1) \rightsquigarrow_i \neg flp$)
- Is a_2 decisive for Alexis?

Decisive argument

Decisive argument

a is *decisive* for i in favor of φ iff it convinces i whenever it appears within the last two arguments:

$$a \hookrightarrow_i \varphi \iff \forall \alpha \mid a \in \alpha_{[\![\#\alpha-1,\#\alpha]\!]} : \alpha \rightsquigarrow_i \varphi$$

Uniqueness

If a is decisive for i in favor of φ, there is no decisive argument for i in favor of any $p \neq \varphi$

Example: decisive argument

- Is a_1 decisive for Olivier? No (not in favor of 0 or $\neg flp$ as $(a_1, a_2) \rightsquigarrow_i flp$ and not in favor of flp as $(a_2, a_1) \rightsquigarrow_i \neg flp$)
- Is a_2 decisive for Alexis? Assuming that $(\dots, a_2) \rightsquigarrow_{\text{Alexis}} flp$ and that $(\dots, a_2, .) \rightsquigarrow_{\text{Alexis}} flp$, it is

Deliberated preference

Deliberated preference

The deliberated preference of i is p iff there is a decisive argument for i in favor of p; if no such $p \in P$ then it is $\emptyset$:

$$\begin{cases} \pi_i = p & \iff \exists a \mid a \hookrightarrow_i p \\ \pi_i = \emptyset & \iff \forall p \in P, \nexists a \mid a \hookrightarrow_i p \end{cases}$$

Example: deliberated preference

- π_{Alexis}?

Deliberated preference

Deliberated preference

The deliberated preference of i is p iff there is a decisive argument for i in favor of p; if no such $p \in P$ then it is $\emptyset$:

$$\begin{cases}\pi_i = p & \iff \exists a \mid a \hookrightarrow_i p \\ \pi_i = \emptyset & \iff \forall p \in P, \nexists a \mid a \hookrightarrow_i p\end{cases}$$

Example: deliberated preference

- π_{Alexis}? *flp*

Outline

At this stage

- Someone's deliberated preference π_i is well defined given $\rightsquigarrow$
- But we don't know $\rightsquigarrow$
- And we can't observe all of it!
- We need to phrase theories and determine how to validate them

Claims

Claim

A claim is a set $C \subseteq P^{\mathscr{A}^{<\mathbb{N}}{}^{I}}$ of attitudes $\rightsquigarrow$ considered as the possible ones

The claim excludes the complementary attitudes!

Example claims

- "Alexis deliberately prefers *flp*" ($C = \{\rightsquigarrow \mid \exists a \mid a \hookrightarrow_{\text{“Alexis}} flp\}$)
- "Olivier never changes his mind given a_1" ($C = \{\rightsquigarrow \mid \forall \alpha : \rightsquigarrow_{\text{Olivier}}(\alpha) = \rightsquigarrow_{\text{Olivier}}(\alpha, a_1)\}$)
- "Olivier reacts exactly like Yves" $[\forall \alpha : \rightsquigarrow_{\text{Olivier}}(\alpha) = \rightsquigarrow_{\text{Yves}}(\alpha)]$
- Combinations of the above

Theories

Claim

A claim is *trivial* iff it contains all attitudes

$$C_{\text{trivial}} = P^{\mathscr{A}^{<\mathbb{N}} I}$$

Theory

A theory is a non trivial claim

The word “theory” should be taken as a technical term here.

Questions to be explored

- What should be postulated about observations? (Observable sets and Anonymity)
- What is a useful theory? (Indicativeness)
- How to ensure the correctness of a theory? (Falsifiability)

Observations

- We cannot "undo" exposure to arguments
- For a given i, we cannot observe both $\rightsquigarrow_i(a_1, a_2)$ and $\rightsquigarrow_i(a_3, a_4)$.
- We can only observe the reactions of i to sets of increasing sequences, such as $\langle(\emptyset), (a_3), (a_3, a_4), (a_3, a_4, a_1), \dots\rangle$

Alexis does not forget

- Assume that we observe that $(a_2) \rightsquigarrow_{\text{Alexis}} \mathit{flp}$
- Now we cannot observe $(a_1) \rightsquigarrow_{\text{Alexis}} \neg \mathit{flp}$
- We can only observe $(a_2, a_1) \rightsquigarrow_{\text{Alexis}} \mathit{flp}$

- However, we can observe incompatible sequences on *different* individuals (e.g. $\rightsquigarrow_i(a_1, a_2)$ and $\rightsquigarrow_j(a_3, a_4)$)

Possible observations

- An observation is a set of triples $\theta \subset \mathscr{A}^{<\mathbb{N}} \times I \times P$
- The possible observations are the finite sets of triples $\theta \subset \mathscr{A}^{<\mathbb{N}} \times I \times P$ such that for a given i, the sequences of arguments related to i in θ forms an increasing sequence
- Let Θ denote that set of possible observations
- Let $\Theta \cap \mathscr{P}(\rightsquigarrow)$ denote the set of possible *observables*: observations that are compatible with $\rightsquigarrow$

Outline

Anonymity

Anonymity requires to not care about the identity of individuals

Anonymous theory

A theory T is anonymous iff it is closed under renaming of individuals:

$$\forall \sigma : I \leftrightarrow I, \rightsquigarrow \in T : (\rightsquigarrow \circ \sigma) \in T.$$

An anonymous theory does not distinguish individuals beyond their attitude as captured by $\rightsquigarrow$ (informational constraint similar to Arrow's IIA).

Anonymity of theories

- "Olivier never changes his mind given a_1"?
- "Everybody opts for the same choice given a_1"?

Anonymity

Anonymity requires to not care about the identity of individuals

Anonymous theory

A theory T is anonymous iff it is closed under renaming of individuals:

$$\forall \sigma : I \leftrightarrow I, \rightsquigarrow \in T : (\rightsquigarrow \circ \sigma) \in T.$$

An anonymous theory does not distinguish individuals beyond their attitude as captured by $\rightsquigarrow$ (informational constraint similar to Arrow's IIA).

Anonymity of theories

- "Olivier never changes his mind given a_1"? Not anonymous
- "Everybody opts for the same choice given a_1"?

Anonymity

Anonymity requires to not care about the identity of individuals

Anonymous theory

A theory T is anonymous iff it is closed under renaming of individuals:

$$\forall \sigma : I \leftrightarrow I, \rightsquigarrow \in T : (\rightsquigarrow \circ \sigma) \in T.$$

An anonymous theory does not distinguish individuals beyond their attitude as captured by $\rightsquigarrow$ (informational constraint similar to Arrow's IIA).

Anonymity of theories

- "Olivier never changes his mind given a_1"? Not anonymous
- "Everybody opts for the same choice given a_1"? Anonymous

Informativeness and indicativeness

- A theory may fail to inform about anyone's deliberated preference (example?)
- A theory may inform only about numbers ("More individuals deliberately prefer *flp* than $\neg$*flp*")
- A theory may indicate something about someone's deliberated preference when knowing some of their reactions to arguments

Indicativeness

A theory T is indicative iff for some observations about i, i's deliberated preference considering any attitude compatible with the observations and T is a strict subset of P

An indicative theory

"If i chooses *flp* given (a_1, a_2) then her deliberated preference is *flp*"

Informativeness and indicativeness

- A theory may fail to inform about anyone's deliberated preference (example? "Olivier never changes his mind given a_1")
- A theory may inform only about numbers ("More individuals deliberately prefer flp than $\neg flp$")
- A theory may indicate something about someone's deliberated preference when knowing some of their reactions to arguments

Indicativeness

A theory T is indicative iff for some observations about i, i's deliberated preference considering any attitude compatible with the observations and T is a strict subset of P

An indicative theory

"If i chooses flp given (a_1, a_2) then her deliberated preference is flp"

Indicativeness

Example (An indicative theory)

"If i chooses flp given (a_1, a_2) then her deliberated preference is flp"

$$[\forall i \in I : (a_1, a_2) \rightsquigarrow_i flp \implies \pi_i = flp]$$

Validity

- So far: syntactic properties (can be checked without querying $\rightsquigarrow$)
- We need to check that the theory *holds*
- Holding is an empirical property

Holding

A theory T holds iff $\rightsquigarrow \in T$

Verifiability

Verifiability

- A theory T is verifiable in principle iff for some observations, T is deducible from the observations

$$\exists \theta \in \Theta \mid \forall \rightsquigarrow \in P^{\mathscr{A}^{<\mathbb{N}^{I}}} : (\theta \subset \rightsquigarrow \implies \rightsquigarrow \in T)$$

- A theory T is verifiable effectively iff for some observables, T is deducible from the observations

$$\exists \theta \in \Theta \cap \mathscr{P}(\rightsquigarrow) \mid \forall \rightsquigarrow \in P^{\mathscr{A}^{<\mathbb{N}^{I}}} : (\theta \subset \rightsquigarrow \implies \rightsquigarrow \in T)$$

Note that effective verifiability ensures that the theory holds. But:

Indicativeness and Verifiability are incompatible

When $\#\mathscr{A} \geq 2$, if T is indicative, then T is not verifiable

Falsifiability: an attempt

Falsifiability (attempt)

A theory T is *falsifiable* iff some observations permits to falsify it:

$$\Theta \nsubseteq \cup_{\rightsquigarrow' \in T} \mathscr{P}(\rightsquigarrow').$$

Fails!

An intuitively non falsifiable theory

- $(a) \rightsquigarrow_i \varphi \vee (a') \rightsquigarrow_i \varphi$ is not falsifiable (okay)
- $\alpha \rightsquigarrow_j \varphi \wedge [(a) \rightsquigarrow_i \varphi \vee (a') \rightsquigarrow_i \varphi]$ is falsifiable (should not be)

Falsifiability

Falsifiability

A theory T is *falsifiable* iff whatever the real attitude is, if it is not in T then we can observe that it is not:

$$\forall{\leadsto} \notin T : \Theta \cap \mathscr{P}(\leadsto) \nsubseteq \cup_{\leadsto' \in T} \mathscr{P}(\leadsto').$$

Falsifiability

- $[\forall i \in I : (a_1) \leadsto_i \mathit{flp}]$?
- Given i: $[(a_1) \leadsto_i \mathit{flp} \vee (a_2) \leadsto_i \mathit{flp}]$?
- $[\exists i \in I \mid (a_1) \leadsto_i \mathit{flp}]$?

Falsifiability

Falsifiability

A theory T is *falsifiable* iff whatever the real attitude is, if it is not in T then we can observe that it is not:

$$\forall \rightsquigarrow \notin T : \Theta \cap \mathscr{P}(\rightsquigarrow) \not\subseteq \cup_{\rightsquigarrow' \in T} \mathscr{P}(\rightsquigarrow').$$

Falsifiability

- $[\forall i \in I : (a_1) \rightsquigarrow_i flp]$? Falsifiable
- Given i: $[(a_1) \rightsquigarrow_i flp \vee (a_2) \rightsquigarrow_i flp]$?
- $[\exists i \in I \mid (a_1) \rightsquigarrow_i flp]$?

Falsifiability

Falsifiability

A theory T is *falsifiable* iff whatever the real attitude is, if it is not in T then we can observe that it is not:

$$\forall \rightsquigarrow \notin T : \Theta \cap \mathscr{P}(\rightsquigarrow) \not\subseteq \cup_{\rightsquigarrow' \in T} \mathscr{P}(\rightsquigarrow').$$

Falsifiability

- $[\forall i \in I : (a_1) \rightsquigarrow_i flp]$? Falsifiable
- Given i: $[(a_1) \rightsquigarrow_i flp \vee (a_2) \rightsquigarrow_i flp]$? Not falsifiable
- $[\exists i \in I \mid (a_1) \rightsquigarrow_i flp]$?

Falsifiability

Falsifiability

A theory T is *falsifiable* iff whatever the real attitude is, if it is not in T then we can observe that it is not:

$$\forall \rightsquigarrow \notin T : \Theta \cap \mathscr{P}(\rightsquigarrow) \not\subseteq \cup_{\rightsquigarrow' \in T} \mathscr{P}(\rightsquigarrow').$$

Falsifiability

- $[\forall i \in I : (a_1) \rightsquigarrow_i \mathit{flp}]$? Falsifiable
- Given i: $[(a_1) \rightsquigarrow_i \mathit{flp} \vee (a_2) \rightsquigarrow_i \mathit{flp}]$? Not falsifiable
- $[\exists i \in I \mid (a_1) \rightsquigarrow_i \mathit{flp}]$? Falsifiable iff I is finite

Satisfiable properties?

- Ongoing work: investigate conditions for simultaneous satisfiability of properties
- For example, it is possible under "reasonable" conditions of regularity to satisfy anonymous and holding together with indicativeness (see below).
- Does there exists attitudes such that no theory that holds is falsifiable and indicative?

Theorem (Sufficient condition for a theory that holds and is anonymous and indicative)

Assume that for some $p \in P$, we have $\exists i \in I \mid P_i = p$ and $\forall i_2 \mid P_{i_2} \neq p, \exists A \in \mathcal{F}(\mathscr{A}) \mid \forall i \in I \mid P_i \neq p, \exists a \in A \mid \hookrightarrow_i(a) \in P \setminus \{p\}$, then there exists a theory that holds and is anonymous and indicative.

Outline

Deliberated preference

- Deliberated preferences complement shallow preferences
- They retain some attractive features about shallow preferences: observability, precision, choice semantics
- Formal definitions about deliberated preferences permit to clarify concepts and compatibilities ("philosophers look for incompatibilities")
- Deliberated preferences could constitute a legitimate basis for individual decision aiding
- Deliberated preferences could constitute a legitimate basis for collective decision making

Normative VS empirical aspects

- Social choice theory separates normative choices (which axioms one wants) from deductive aspects (which are compatible; what rule to use)
- This endeavor: separate the normative choice (the set of arguments, the protocol of observation, the desired properties of theories) from the empirical content (which theories are valid, which arguments convince individuals)
- This approach may permit to frame some disagreements about what to do as empirical questions
- Long term goal: study sophisticated opinionated normative theories (Rawls, Nozick, Chomsky); useful for studying nudging

Thank you for your attention!