Departures from Preference Maximization and Violations of the Sure-Thing Principle

Geoffroy de Clippel

June 2023[*]

Abstract

A choice-based version of the sure-thing principle is most often violated without preference maximization. Implications for (individual and social) choice theory, game theory and mechanism design include: dynamic consistency with respect to the resolution of uncertainty implies rationality over constant acts, totally-mixed beliefs cannot be overlooked when checking dominance in games, dominance in extensive-form games is not equivalent to dominance when choosing behind the veil of ignorance in their associated strategic forms, generalizations of serial dictatorship, and only those rules, are dominant-strategy implementable over the largest domain of all choice functions, and it becomes preferable to use dynamic mechanisms.

1 Introduction

Savage (1972, page 39) motivates the sure-thing principle with the following story:

> *A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant to the attractiveness of the purchase. So, to clarify the matter for himself, he asks whether he would buy if he knew that the Republican candidate were going to win, and decides that he would do so. Similarly, he considers whether he would buy if he knew that the Democratic candidate were going to win, and again finds that he would do so. Seeing that he would buy in either event, he decides that he should buy,*

[*] The paper is the first complete version of a preliminary draft entitled '*Incorporating Risk in Choice Theory: Some Observations*' that circulated in 2020. I thank Shengwu Li, Pietro Ortoleva, Marek Pycia and Kareen Rozen, as well as seminar participants at the Paris School of Economics and Princeton University, for useful comments.

> *even though he does not know which event obtains, or will obtain, as we would ordinarily say.*

The story describes what *choices* the businessman would make in different states of the world, and concludes that he should *choose* to buy when not knowing which event will obtain if he would do so in either one of them. Of course, Savage could have equivalently described the story in terms of *preferences* since choices are derived through preference maximization in his framework. But the story does sound reasonable enough expressed in choices as it is, independently of how these choices arise and what other options the businessman has.

The present paper formulates a choice-based property in this spirit, see STP below. Under rationality, the property is mild and hence most often taken as granted. It is satisfied in particular whenever maximizing a preference consistent with first-order stochastic dominance (including for instance expected-utility or rank-dependent utility, etc.).[1] Things are different when considering irrational choice functions, as the two following examples illustrate.

Consider first a shortlisting method in the spirit of Manzini and Mariotti (2007) to capture individual choices that need not be rational.

Example 1. *When selecting a wine bottle, an individual first looks at its vintage and region of production. He starts by eliminating any bottle that is Pareto inferior to another available option along both dimensions. He then examines closely the surviving options, weighing carefully their pros and cons, to finally reach his final choice. Say this last step is consistent with maximizing a standard preference relation. The table below provides the individual's utility (u), and ratings along the two selection criteria (sc_1 and sc_2), over three bottles (x, y, and z).*

	sc_1	sc_2	u
x	1	4	2
y	2	2	3
z	3	3	1

For instance, the individual would end up selecting y from $\{x, y\}$, as both options remain on the shortlist and y provides a larger utility, and z from $\{y, z\}$ because y is Pareto inferior to z for the preliminary selection criteria.

While one could contemplate other avenues,[2] suppose the individual applies a similar procedure, this time to expected values, when risk is involved. For instance, he would end up

[1] Prospect theory was criticized for not agreeing with first-order stochastic dominance, and adapted in a 'second wave' of models to fix what was viewed as a flaw. More recent versions with endogenous reference points, on the other hand, can violate the property.

[2] We develop the example around one possible, arguably natural extension of the shortlisting procedure to lotteries. This is solely for illustration purposes, as the main result below applies *whatever* the extension.

selecting $\ell = 0.5x \oplus 0.5y$ over $\ell' = 0.5y \oplus 0.5z$, because both lotteries remain on the shortlist[3] *and $u(\ell) = 2.5 > 2 = u(\ell')$.*

Suppose now a card will be drawn from a deck of Black and Red cards. The individual must choose between two bets (or 'acts'):

	Black	*Red*
Bet 1	x	y
Bet 2	y	z

The individual picks Bet 2 both if he knows the state is Black (as it gives y while choosing between bottles x and y), or if he knows the state is Red (as it gives z while choosing between bottles y and z). Since he picks Bet 2 whatever the card color, one may expect him, in the spirit of the sure-thing principle, to pick that same bet whatever the proportion p of black cards in the deck. This is wrong, as it follows from the previous paragraph that he picks Bet 1 when $p = 1/2$ (and, more generally, whenever p is strictly between $1/3$ and $2/3$).

Consider now group decisions. This time, all participants are rational, but their aggregate choices fail to be rationalizable.

Example 2. *A committee applies the Borda rule to reach decisions. Its members' utilities for three possible outcomes (x, y or z) are:*

	u_1	u_2	u_3
x	2	1	3
y	4	2	0
z	0	4	2

Say there are n_i committee members with utility function u_i. With $n_i + n_j > n_k$ for all distinct $i, j, k \in \{1, 2, 3\}$, we get the oft-discussed Condorcet cycle, with y selected from $\{x, y\}$, z selected from $\{y, z\}$, and x selected from $\{x, z\}$.

Assuming committee members are expected utility maximizers,[4] *the Borda rule also applies when selecting lotteries. For instance, the committee selects $\ell = 0.5x \oplus 0.5y$ over $\ell' = 0.5y \oplus 0.5z$, because $u_i(\ell) > u_i(\ell')$ for $i = 1, 3$.*

Suppose now the committee must select one of two courses of actions (a or b), whose outcome is impacted by a state (ω_1 or ω_2):

	ω_1	ω_2
a	x	y
b	y	z

[3] $sc_1(\ell) = 1.5 < 2.5 = sc_1(\ell')$ and $sc_2(\ell) = 3 > 2.5 = sc_2(\ell')$.

[4] A similar comment as in footnote 2 applies here as well.

With y (resp., z) being the committee's selection from $\{x, y\}$ (resp., $\{y, z\}$), it selects b both if its members know the state is ω_1, and if they know the state is ω_2. Yet, against the spirit of the sure-thing principle, the committee selects a when its members think ω_1 and ω_2 are equally likely to obtain (and, more generally, whenever p is strictly between $2/5$ and $2/3$), as follows from the previous paragraph.

These are but two examples. Are these violations of the sure-thing principle a peculiarity of the shortlisting method/Borda rule and their specific extensions to the domain of lotteries? In other words, how common are such violations of the sure-thing principle beyond preference maximization? They are very prevalent, as it turns out (see Proposition 1): *any choice function satisfying STP must be rational over deterministic outcomes/constant acts*. Think now of choice theory's revival over the past fifteen years or so (see de Clippel and Rozen (forthcoming) for a survey), including for instance Manzini and Mariotti (2007) mentioned for Example 1. Most papers in this literature study choice processes individuals may use to make decisions, and choice functions they trigger, over a finite set O of options, so as to accommodate behavioral biases that are inconsistent with rationality. Our result implies that *any* extension of these choice functions capturing irrational behavioral biases to menus of lotteries (subsets of $\Delta(O)$) must violate STP. Alternatively, we know from Arrow's impossibility theorem that irrationality oftentimes prevails in group choices when O contains at least three elements, even when group members are rational. Hence some STP violations *must occur* when these groups face risk. The result also means that any choice function violating rationality over constant acts must display some dynamic inconsistency with respect to the resolution of uncertainty.

Beyond (individual and social) choice theory, these observations have implications in game theory and mechanism design. Under preference maximization, a player's strategy is dominant if it is a best response whatever its belief about its opponents' actions. But optimality is typically checked only against opponents' *pure* strategies. This is fine when restricting attention to expected utility (or any preference ordering satisfying first-order stochastic dominance). The first lesson is that one should not overlook probabilistic beliefs when checking dominance in the absence of preference maximization. This makes dominance harder to achieve than one might think. The second lesson, this time for dynamic games, is that checking dominance at the moment a player makes her choice, as it should be, can be more permissive than checking dominance of complete strategies before the game starts, as one is used to do (and correctly so in standard models). In particular, in sharp contrast to the expected-utility benchmark, the equivalence of extensive-form games and their associated strategic forms for identifying dominant strategies may fail.

To summarize, STP violations prevail when deterministic choice functions fail to be rational. As a result, dominance becomes even more demanding than one is used to, though less severly so in dynamic games. One is then left wondering: Are we facing essentially an impossibility result, establishing we should abandon the notion of dominant strategy beyond preference maximization? Or, is there hope the concept brings valuable insights in relevant applications? As a first step in addressing those questions, Section 4 covers a mechanism-design application, a field where dominant strategies play an important role. Specifically, we tackle problems of unit-demand assignment problems, one of the building blocks of the market-design literature. A participant's type must now capture her choice function over lotteries of outcomes (instead of merely a preference over deterministic outcomes in the standard, rational benchmark): though mechanisms are deterministic, STP violations mean one cannot overlook totally-mixed beliefs about others' strategies. These choice functions can be fully rational, with expected or non-expected utility, or capture more complex choice patterns reflecting behavioral biases and STP violations. Needless to say, implementation over this richest domain is most challenging, but also extremely desirable given its robustness to all forms of behaviors.

Proposition 2 establishes that the class of all non-bossy social choice functions that are implementable in dominant strategies can be represented as generalized serial dictatorships. Though perhaps intuitive enough ex-post, it wasn't obvious at the outset that any rule would be robustly implementable this way, and proving necessity (that no other rule fits the bill) is by no means trivial. Also, identifying these rules is possible only after recognizing that dynamic games are oftentimes preferable in this context (as highlighted in Section 3), because observing past choices reduces the uncertainty players face. Indeed, implementing generalized serial-dictatorship rules often *requires* a dynamic mechanism, in sharp contrast to dominant-strategy implementation over the rational domain (where restricting attention to static mechanisms is without loss).

To conclude, one should exercise caution, and avoid misleading intuition gleaned from experience with more standard frameworks when incorporating risky prospects in models of bounded rationality or group choices. But important lessons can be learned when doing it properly.

Related Literature

A recent literature highlights people's difficulty in performing contingent reasoning. While agents are assumed to be payoff maximizing, such mistakes are attributed to the complexity of projecting oneself in multiple mutually-exclusive circumstances. Esponda and Vespa (2021) is perhaps most related within that literature given its focus on Savage's sure-thing principle.[5] They revisit five classic anomalies in decision and game theory under two frames. While the first one present these problems in their standard format, the second frame is designed to make contingent thinking more salient. Their data provides evidence that these anomalies are partly driven by a failure of the sure-thing principle. Our theoretical contribution also speaks to violations of a sure-thing principle, but for an orthogonal reason: even if one maintains the classic assumption that people can perform contingent reasoning, violations of STP (a *choice-based* analogue of the sure-thing principle) arise when choices over deterministic outcomes cannot be explained through preference maximization. The sections on mechanism design and behavioral implementation in dominant strategies become only more relevant if further STP violations arise because some people have trouble performing contingent reasoning.

Li (2017) introduces obvious dominance as a more restrictive alternative to standard dominance in games. A player's strategy s is *obviously dominant* if, for all deviations, the best payoff she obtains under all possible ensuing outcomes is smaller than the worst payoff she obtains under all outcomes that may arise when sticking to s.[6] Pycia and Troyan (forthcoming) introduce a stronger version of obvious dominance in their study of simplicity in games and mechanisms, whereby a player at an information set treats her own future moves as moves from another party. Either way, while players are payoff maximizing, dominance is tested with inconsistent beliefs (most pessimistic for s and most optimistic under any deviation). By contrast, participants (individuals or groups) in the present paper need not be payoff maximizers, and dominance is simply the choice-based extension of the standard definition (where beliefs remain unchanged when performing comparisons). Though for different reasons, a common consequence is that finding a dominant strategy in the mechanism-design problem becomes harder under both approaches, which narrows the set of implementable rules. Also, both approaches prove the value of dynamic games in contrast to the standard framework where it is sufficient to study static mechanisms when it comes to dominant-strategy implementation.

Following Pápai (2000), Pycia and Ünver (2017) fully characterize the set of social choice functions that are efficient, non-bossy and implementable in dominant strategies for unit-demand assignment problems.[7] A subset of them, Pápai's hierarchical exchange rules satisfying a dual-ownership property, are furthermore implementable in obviously dominant

[5] The reader who is interested in failures of contingent reasoning beyond violations of the sure-thing principle, is referred to Esponda and Vespa (2021)'s discussion of the related literature.

[6] These comparisons are made at the time the deviation starts, which is important to determine what set of outcomes must be contemplated.

[7] With unit demand, group strategyproofness is equivalent to dominant-strategy implementation plus non-bossiness on the rational domain. We won't focus on group strategyproofness as it is not immediately clear how to define it beyond the rational domain.

strategies, see Troyan (2019) and Mandal and Roy (2022). Using strong obvious dominance, Pycia and Troyan (forthcoming, Section 4.3) further narrows down the set of implementable rules, singling out essentially the same mechanisms as we do.[8] These works all assume participants are payoff maximizers, and determine dominance by comparing deterministic outcomes arising under the mechanism (using consistent or inconsistent deterministic beliefs about what happens in the game on path and after a deviation). By contrast, we permit a much larger class of behaviors, indeed allowing for any choice function over lotteries, while observing that dominance should hold given all (consistent) stochastic beliefs about others' choices. Thus, being dominant must hold both for the deterministic outcomes arising when the player believes others play pure strategies, and for lotteries arising when the player has totally-mixed beliefs. While the former implies the latter under the sure-thing principle, our arguments highlight that being dominant becomes very demanding for choice functions that violate STP (which, as shown in the first part of the paper, must occur when allowing for non-rational choice functions over constant acts). Yet implementation in dominant strategy remains doable. Characterization results (with or without non-bossiness, and with or without Pareto efficiency on the rational domain) gravitate around generalized serial dictatorship rules. Known to be implementable in (strongly) obviously-dominant strategies, these rules also have the remarkable feature of guaranteeing the existence of dominant strategies whatever the participants' choice functions. This is important not only because of the many behavioral biases that have been documented in the literature on individual choices, but also because the aggregation of multiple conflicting preferences may fail to be rational when participants are groups (e.g., assigning public housing to families).

As detailed in Dreyfuss, Heffetz and Rabin (2022), recent evidence, both empirical and experimental, shows a substantial fraction of dominated choices in the widely-implemented deferred acceptance algorithm. Interestingly, they prove this behavior may be partly intentional under expectations-based loss aversion (as in Köszegi and Rabin (2009)). Their argument centers around a key property of these non-expected utility preferences, namely that they violate first-order stochastic dominance (and hence STP). They also show that reasonable specifications of their behavioral model better fits Li (2017)'s experimental data on random serial dictatorship[9] than standard expected utility. Dreyfuss, Glicksohn, Hef-

[8] Comparing these last three papers, it is interesting how the class of implementable social choice functions is further trimmed when imposing the strong form of obvious dominance. In other words, it does matter under this approach whether or not a player knows, and sticks to what her own strategy prescribes at future information sets. By contrast, a player is never uncertain about her own future moves when comparing strategies under the consistent-beliefs approach we pursue. What needs to be eliminated is the uncertainty arising from others' future choices when accommodating all choice functions.

[9] Players submit a ranking, independently of each others and knowing only their own priority score. They face thus much uncertainty when choosing a report, which can break down the dominance of strategies in

fetz, and Romn (2022) experimentally test four DA variants – {static, dynamic} × {student proposing, student receiving} – and show that, while predicted behavior is identical in all four formats under standard preferences, incorporating expectations-based reference dependence predicts important differences. New experimental data they collect confirms these differences. In particular, the dynamic student-receiving version leads to the highest compliance with straightforward behavior. Here too we see that restricting attention to static mechanisms is with loss of generality.

The paper also contributes to the growing effort of incorporating lessons from behavioral economics into the implementation and mechanism-design literatures. The closest contribution is de Clippel (2014) who investigates behavioral Nash implementation under complete information while allowing participants' choice to diverge from preference maximization. The case of dominant strategies under private information is equally important. Sections 2 and 3 provide the groundwork for understanding what dominance means and entails in this context, while Section 4 then pursues the implementation exercise itself for a relevant class of problems.

2 Starting Point

Let O be a (finite) set of relevant *outcomes*. A *lottery* is a probability distribution over O. A *choice function* c associates to each (finite) set L of lotteries a subset $c(L)$. Since outcomes are degenerate lotteries, a choice function also defines choices over subsets of O. The restriction of c on subsets of O is assumed to be single-valued. This restriction is *rational* if there exists a preference ordering $\succ$ on O such that $c(S) = \arg\max_{\succ} S$, for each $S \subseteq O$.

Let Ω be a (finite) set of *states of the world*. An *act* is a map that associates an outcome to each state of the world. The states' relative likelihoods are captured by a probability distribution $p \in \Delta(\Omega)$. Assuming state-independence, as we do throughout the paper, means that only lotteries associated to acts matter to the decision-maker. Given any finite set A of acts, let $L^p(A)$ be the set of lotteries $\ell^p(a)$ – "outcome x occurs with probability $\sum_{\omega|a(\omega)=x} p(\omega)$" – obtained by varying $a \in A$. If p puts probability one on ω, then $L^p(A)$ is denoted $A(\omega)$, that is, $A(\omega) = \{a(\omega)|a \in A\}$.

In the spirit of the sure-thing principle, we investigate the following property on choice functions over risky prospects:

Property STP *Let A be a set of acts, and $a \in A$. If $c(A(\omega)) = \{a(\omega)\}$ for each $\omega \in \Omega$, then $c(L^p(A)) = \{\ell^p(a)\}$ for all $p \in \Delta(\Omega)$.*

the presence of STP violations.

If the decision-maker knows that the state is ω, then picking an act amounts to choosing an outcome within $A(\omega)$. Now suppose that the act a has the unique property of delivering her chosen outcome in *each* state ω. This is a very stringent property that often does not apply as an act may provide the chosen outcome in some state but rarely in all states. Given state-independence, picking an act from A given p amounts to picking a lottery from $L^p(A)$. STP requires this lottery to be $\ell^p(a)$, indeed the one associated to act a. Inspired by the sure-thing principle, this seems reasonable at first because a delivers the outcome she wants to pick whatever the state realization. Yet, as claimed in the Introduction, this property is violated as soon as the restriction of c to deterministic outcomes is not rational.

Proposition 1. *If c satisfies STP, then it is rational over deterministic outcomes.*

The two examples from the introduction leverages the specifics of the Borda rule and the shortlisting method. Proposition 1 reveals that these specifics do not matter for establishing the existence of STP violations: the mere presence of IIA violations over deterministic outcomes suffices.[10]

Remark 1. *Are there irrational choice functions satisfying STP while consistent with preference maximization over deterministic outcomes? For concreteness, suppose that outcomes are monetary payoffs and that choices are obtained by payoff maximization in the absence of risk. Say the choice function c is* consistent with first-order stochastic dominance *if, for each set of lotteries L over monetary amounts, there is no lottery $\ell \in L$ that first-order stochastically strictly dominates $c(L)$. As is easily checked, any such choice function satisfies STP.*[11] *In particular, one can construct many irrational choice functions over monetary lotteries satisfying STP: for any choice function c, the modified choice function $\hat{c}$ that selects, from any menu of monetary lotteries the choice from the subset of lotteries that are not first-order strictly dominated, satisfies STP.*

[10] IIA violations occur in both examples given that x is selected from $\{x, y, z\}$, but y is selected from $\{x, y\}$.

[11] Other choice functions may violate STP. For an example, consider a "*cautious investor comparing alternative portfolios [who] first eliminates those that are too risky relative to others available, and then ranks the surviving ones on the basis of expected returns*" (Manzini and Mariotti (2007, page 1825)). Concretely, suppose the investor first computes each lottery's coefficient of variation, and eliminates those with a coefficient strictly above average (within the set of available lotteries). While rational over deterministic outcomes, the investor violates STP by selecting an asset paying $\$1,500$ whatever the state, over one paying $\$1,501$ if the state belongs to $\Omega' \subset \Omega$ and $\$1,502$ otherwise, whenever she places strictly positive probability on both Ω' and $\Omega \setminus \Omega'$. Expectations-based loss aversion (see e.g. Köszegi and Rabin (2009)) provides another example.

3 Dominant Strategies in Game Theory

3.1 Lesson for strategic-form games

A *rational strategic-form game* specifies for each player $i \in N$ a finite set S_i of strategies, a preference ordering $\succ_i$ over outcomes in O,[12] and an outcome function $f : S \to O$ where $S = \times_{i \in N} S_i$ is the set of strategy profiles. In many courses and textbooks, the prisoners' dilemma is offered as a first illustration, immediately followed by the notion of dominant strategy (see e.g. the exposition in Mas-Colell, Whinston and Green (1995, Section 8.B)). Formally, strategy s_i^* is *dominant* for player i if $f(s_i^*, s_{-i}) \succ_i f(s)$ for all $s \in S$ such that $f(s_i^*, s_{-i}) \neq f(s)$. Later on, one usually points out that a player's belief about their opponents' strategies need not be deterministic (either because they believe others randomize, or because they are unsure about which strategy others select). One must now define each player i's preference over lotteries. For this, one typically specifies i's Bernoulli utility for each outcome (that agrees with $\succ_i$, while also capturing risk attitudes), and assumes strategies are selected by maximizing expected utilities. Contrary to the notions of dominated strategy or Nash equilibria, the definition of a dominant strategy is usually not revisited at that point. This is sensible since expected utility guarantees that a dominant strategy remains superior to all alternative strategies, *whatever the player's belief about his opponents' strategies.*

The first lesson in this section warns that, by contrast, considering stochastic beliefs ceases to be optional beyond the rational domain, because of systematic violations of STP. To formalize this simple, but important observation, consider now players using choice functions that need not be compatible with preference maximization (either because of behavioral biases or because players are groups instead of individuals). The notion of a game easily extends: a *behavioral strategic-form game* is obtained simply by substituting in the definition each player i's preference over lotteries (expected utility for a prespecified Bernoulli utility consistent with $\succ_i$) by her choice function c_i.

Definition 1. *Strategy s_i^* is* dominant *for i* against deterministic strategies *if*

$$c_i(\{f(s) | s_i \in S_i\}) = \{f(s_i^*, s_{-i})\},$$

for all $s_{-i} \in S_{-i}$.

If i expects others to pick s_{-i}, then the opportunity set of outcomes she faces when picking her own strategy is $\{f(s) | s_i \in S_i\}$. Suppose she'd pick o from that set, which

[12] We leave aside the possibility of indifference over deterministic outcomes in line with our focus on single-valued choice functions on that domain.

happens to be precisely the outcome she gets when picking s_i^*. For s_i^* to be dominant against pure strategies, this property must hold whatever s_{-i}. This is the natural extension of the property of dominance one checks in rational games.

But dominance requires the strategy to be selected whatever the player's belief about her opponents' strategies. To define this, consider now c_i defined over $\Delta(O)$. Given a belief $p \in \Delta(S_{-i})$, let $\ell^p(s_i)$ be the lottery that selects $f(s)$ with probability $p(s_{-i})$ and let $L_i^p = \{\ell^p(s_i) | s_i \in S_i\}$ be the opportunity set of lotteries that i faces when picking her strategy.

Definition 2. *Strategy s_i^* is* dominant *for i if $c_i(L_i^p) = \{\ell^p(s_i^*)\}$ for all $p \in \Delta(S_{-i})$.*[13]

Lesson 1. *While Definitions 1 and 2 are equivalent in standard*[14] *rational games, this equivalence fails to extend to the class of behavioral strategic-form games.*

This follows at once from our initial observation in Section 2 after realizing that other players' strategies can play the roles of states. The equivalence in standard rational games holds because optimality against each pure strategy guarantees optimality against opponents' correlated strategies when STP holds. By contrast, consider the committee from Example 2 picking a row in the following game:

	L	R
a	x	y
b	y	z

Earlier arguments now mean that b is dominant for the committee against its opponent's pure strategies, while failing to be dominant, as the committee would pick a if its members view 'L' and 'R' are equally likely (or occuring with a probability strictly between 2/5 and 2/3). Beyond this example, it is straightforward to replicate the proof of Proposition 1 to show that, for each choice function c failing rationality over O, there exists a behavioral strategic-form game with outcomes in O for which a player who follows c has a strategy failing to be dominant while being dominant against pure strategies.

[13] Under this definition, a player's belief allows for correlation across opponents' strategies. This is more demanding than requiring dominance for all independent beliefs, that is, for all $p \in \times_{j \neq i} \Delta(S_j)$. As argued by Brandenburger and Dekel (1987), both views are possible. In the setting of a controlled lab experiment, without any coordination device and communication opportunities across players, independence seems reasonable. Otherwise it seems one should entertain the larger class of beliefs. We follow the more demanding, and robust approach.

[14] Meaning that i's preference over lotteries is FOSD-consistent with $\succ_i$ (e.g., expected utility extension).

3.2 Lesson for extensive-form games

Another lesson arising from Proposition 1 pertains to dominant strategies in the richer framework of dynamic games.

An *extensive-form game* is a tree (that is, a rooted graph with no cycle) with a player attached to each non-terminal node and an outcome in a set O attached to each terminal node. Let n be a decision node, and let i be the associated player. The set of arcs getting out of a node is the set of actions available to the player at that node. To model the possible lack of observation about other players' past moves, an *information set* for a player i in an extensive-form game is a collection of decision nodes such that

- player i is attached to every node in the information set,
- the set of actions is identical at each node in the information set.

An extensive-form game is of *perfect information* if all the information sets are singletons. We assume there is *perfect recall*, meaning that players remember their own past moves each time they make a decision. Under the rational benchmark, players are endowed with a preference ordering over O, which is extended to a preference over $\Delta(O)$ in a way that is FOSD-consistent (e.g., expected utility). We also consider the behavioral generalization where each player is endowed with a choice function defined over menus in $\Delta(O)$. A *strategy* for a player in the extensive form is a complete plan of action, that is, the selection of an action at each node she is attached to. The *associated strategic form* is obtained by having each player pick a strategy independently of each others before the start of the game, with the outcome function simply selecting for each strategy profile the outcome attached to the terminal node when implementing those strategies in the extensive form.

The standard approach, developed for the rational benchmark, is to call a strategy dominant in the extensive-form if it is dominant in its associated strategic form. In other words, it does not matter whether dominance is assessed at the time a player actually makes her choice, or before the game starts. Aligning with the meaning of an extensive form (whose purpose is to capture the sequence of moves and, more importantly, what information players have about past moves when they make their own decisions), only the former scenario makes sense. But of course this distinction is moot under the rational benchmark: by the sure-thing principle, it is equivalent to take the approach behind the veil of ignorance about past actions, akin to implementing a strategy-method approach in experiments, because dominance at each future information set implies dominance whatever the player's belief about the likelihood of reaching those sets. The second lesson following Proposition 1 is that such equivalence does not extend to the class of behavioral extensive-form games because of STP violations.

To formalize this, consider one of player i's information sets, $\mathcal{I}$. The strategy profile s_{-i} for the other players is *consistent with* $\mathcal{I}$ if one can find a strategy s_i for i such that $\mathcal{I}$ is reached when implementing s_i and s_{-i}. Let $S_{-i}^{\mathcal{I}}$ denote the set of other players' strategy profiles that are consistent with $\mathcal{I}$. An information set $\mathcal{I}'$ *succeeds* an information set $\mathcal{I}$ if there is a path from the root to a node in $\mathcal{I}'$ that passes through $\mathcal{I}$. A *strategy for i starting at* $\mathcal{I}$, $s_i^{\mathcal{I}}$, pins down an action at $\mathcal{I}$ and all her subsequent information sets. Let $S_i^{\mathcal{I}}$ denote the set of all such strategies. Conditional on having reached $\mathcal{I}$, the outcome when she follows $s_i^{\mathcal{I}}$ while others follow $s_{-i} \in S_{-i}^{\mathcal{I}}$ is denoted $o(s_i^{\mathcal{I}}, s_{-i})$.[15] Not knowing which strategy profile others follow in that set, all beliefs should be considered. If her belief is $\mu_i^{\mathcal{I}} \in \Delta(S_{-i}^{\mathcal{I}})$, then playing $s_i^{\mathcal{I}}$ gives rise to the lottery $\ell_i^{\mathcal{I}}(s_i^{\mathcal{I}}, \mu_i^{\mathcal{I}})$ where outcome $o \in O$ occurs with probability

$$\sum_{s_{-i} \in S_{-i}^{\mathcal{I}} | o(s_i^{\mathcal{I}}, s_{-i}) = o} \mu_i^{\mathcal{I}}(s_{-i}).$$

Agent i's opportunity set of lotteries when varying her own strategy at $\mathcal{I}$, and given her belief $\mu_i^{\mathcal{I}}$, is thus:

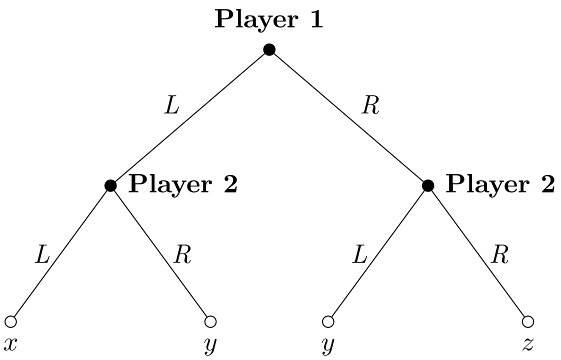
$$\mathcal{O}_i^{\mathcal{I}}(\mu_i^{\mathcal{I}}) = \{\ell_i^{\mathcal{I}}(s_i^{\mathcal{I}}, \mu_i^{\mathcal{I}}) | s_i^{\mathcal{I}} \in S_i^{\mathcal{I}}\}.$$

Playing $s_i^{\mathcal{I}}$ is *dominant for i at* $\mathcal{I}$ if $c_i(\mathcal{O}_i^{\mathcal{I}}(\mu_i^{\mathcal{I}})) = \{\ell_i^{\mathcal{I}}(s_i^{\mathcal{I}}, \mu_i^{\mathcal{I}})\}$, for all $\mu_i^{\mathcal{I}} \in \Delta(S_{-i}^{\mathcal{I}})$. That is, whatever her belief about others' strategies (consistent with $\mathcal{I}$), $s_i^{\mathcal{I}}$ provides the lottery she wishes to choose within the opportunity set $\mathcal{O}_i^{\mathcal{I}}(\mu_i^{\mathcal{I}})$ of lotteries she can obtain by picking a strategy at $\mathcal{I}$. Player i's strategy s_i is *dominant* for i if, for each information set $\mathcal{I}$ at which i plays, $s_i^{\mathcal{I}}$ (the restriction of s_i to $\mathcal{I}$ and subsequent information sets) is dominant for i at $\mathcal{I}$.

Lesson 2. *The strategy in a rational extensive-form game is dominant if, and only if, it is dominant in the associated strategic form. This equivalence fails to extend to the class of behavioral extensive-form games.*

The equivalence under rationality follows from STP, as already explained. To see how issues arise when dropping preference maximization, consider the following perfect-information game:

[15] A unique outcome is determined, because reaching $\mathcal{I}$ pins down i's action choice at all her information sets preceding $\mathcal{I}$ (independently of s_{-i}). Suppose, on the contrary, that one can find an earlier information set $\mathcal{I}'$ and two actions a and a' on different paths going from the root to $\mathcal{I}$. A cycle would arise in the graph if both paths reach the same node in $\mathcal{I}$, contradicting the definition of an extensive-form game as a tree. But reaching two distinct nodes in $\mathcal{I}$ now contradicts perfect recall.



Its associated strategic form is:

	LL	LR	RL	RR
L	x	x	y	y
R	y	z	y	z

Suppose now Player 2 is the committee from Example 2 with, in addition, $3n_2 < n_1 + n_3$ (take for instance $n_1 = n_3 = 2$ and $n_2 = 1$). Though it does not have a dominant strategy in the associated normal form,[16] RR is clearly dominant for the committee in the extensive form because it picks y from $\{x, y\}$ and z from $\{y, z\}$.

4 Illustration: A Case of Behavioral Implementation in Dominant Strategies

To further illustrate the implications of our initial observation in Section 2, and apply in particular the game-theoretic lessons from the previous section, we now revisit the classic problem of dominant-strategy implementation in a behavioral framework, that is, with types determining choice functions instead of preferences (or, equivalently, choice functions satisfying IIA). As we saw, we cannot rely on the sure-thing principle anymore. In particular, a proper definition of strategyproofness must now include the possibility of stochastic beliefs, and dynamic games may become valuable (as equivalence with dominance in the associated normal form breaks down).

Implementation in dominant strategies is already demanding on the rational domain. Further expanding the domain makes it only harder. In view of Gibbard (1973) and Sat-

[16] As discussed in Example 2, the committee selects y (resp., z) from $\{x, y\}$ (resp., $\{y, z\}$), and hence can only select RL or RR (resp., LR or RR) when all members believe 1 picks L (resp., R). With this, RR is the only candidate for a dominant strategy, but LL gets a larger Borda count when $3n_2 < n_1 + n_3$ and 1 is equally likely to pick L or R.

terthwaite (1975), possibility results are typically attained by restricting the structure of the problem and/or the class of possible preferences. Though the literature has successfully explored other important domains (see Barberà (2011)), the present section is dedicated to the study of discrete allocation problems with unit demand. Focusing on this framework is both for convenience (an immediate[17] extension of the discrete-choice model studied thus far) and because of its importance (a main building block of the market-design literature). These problems have been studied extensively on the rational domain, see e.g. the important contributions of Pápai (2000) and Pycia and Ünver (2017).

A set N of *agents* have unit demand over a set X of *indivisible items*.[18] Accommodating the possibility that some agents might not consume any item, let $X^* = X \cup \{\emptyset\}$. An *assignment* α allocates up to one item per agent: $\alpha(i) \in X^*$ with the feasibility constraint that two distinct agents must be assigned distinct items ($\alpha(i) = \alpha(j) \neq \emptyset$ implies that $i = j$). The set of all assignment profiles is denoted $\mathcal{A}$. A *social choice function* f selects an assignment for each choice function profile, that is, $f : \Theta^N \to \mathcal{A}$. Agents are assumed to know their own choice function, but not that of others. A *game form* is an (finite) extensive-form game with agents as players, assignments as outcomes, and preferences/choice functions left unspecified.

Under the standard approach, one would introduce at this point a domain of preferences over X^*, and observe that appending a preference profile to the game form defines an extensive-form game. The novelty of our approach lies in the recognition that agents may fail to be rational, either because of individual behavioral biases or because of challenges when aggregating diverging opinions. As illustrated in the past two sections, it is not sufficient to define choices for menus of deterministic outcomes, because non-degenerate beliefs about others' actions in the game form generate lotteries over X^* (even though the game form determines deterministic assignments), and STP violations imply that choices over resulting menus of lotteries cannot be inferred from choices over menus of deterministic outcomes associated to degenerate beliefs. Instead, we must define, for each agent, a choice function c defined over $\Delta(X^*)$. As before, we assume that c selects from each nonempty menu $S \subseteq X^*$ a single element $c(S) \in S$.[19] Let Θ represent the set of all such choice functions: c_θ is distinct from $c_{\theta'}$, for each $\theta \neq \theta'$, and for each choice function c there is $\theta \in \Theta$ such that $c_\theta = c$. A *choice function profile* specifies a choice function for each agent. The set of all choice

[17] By contrast, studying quasi-linear problems, or exchange with divisible goods, would require to first start with non-rational individual choice functions on those domains. The paper focuses on discrete-choice problems instead because they are more often the focus of the literature on behavioral choice theory.

[18] Agents could be individuals or groups of people (e.g. family), depending on context.

[19] Similarly, preferences are most often assumed to be strict when analyzing unit-demand assignment problems in the rational benchmark.

function profiles is thus Θ^N. Appending a choice function profile θ to a game form defines a behavioral extensive-form game.

Notice that outcomes in this applications are assignments (n-vectors with components in X^*), and agents care only about their assigned item, which is why choice functions are naturally defined over subsets of X^* instead of subsets of $\mathcal{A}$. To reflect this, we denote by $o_i(\sigma_i^{\mathcal{I}}, s_{-i})$ the element of X^* assigned to i conditional on having reached $\mathcal{I}$, when she follows $s_i^{\mathcal{I}}$ while others follow $s_{-i} \in S_{-i}^{\mathcal{I}}$. Given a belief $\mu_i^{\mathcal{I}} \in \Delta(S_{-i}^{\mathcal{I}})$, playing $s_i^{\mathcal{I}}$ gives rise to the lottery $\ell_i^{\mathcal{I}}(s_i^{\mathcal{I}}, \mu_i^{\mathcal{I}})$ over $\Delta(X^*)$ where, for each $x \in X^*$, i gets x with probability

$$\sum_{s_{-i} \in S_{-i}^{\mathcal{I}} | o_i(s_i^{\mathcal{I}}, s_{-i}) = x} \mu_i^{\mathcal{I}}(s_{-i}).$$

Thus agent i's opportunity set of lotteries when varying her own strategy at $\mathcal{I}$, and given her belief $\mu_i^{\mathcal{I}}$, is viewed as a subset of $\Delta(X^*)$ instead of a subset of lotteries over assignments:

$$\mathcal{O}_i^{\mathcal{I}}(\mu_i^{\mathcal{I}}) = \{\ell_i^{\mathcal{I}}(s_i^{\mathcal{I}}, \mu_i^{\mathcal{I}}) | s_i^{\mathcal{I}} \in S_i^{\mathcal{I}}\}.$$

All other definitions (in particular that of a dominant strategy) and observations from Section 3.2 now carry through unchanged.

A game form *implements* the social choice function f *in dominant strategies* if there are functions $(s_i^* : \Theta \to S_i)_{i \in N}$ such that, for each $(\theta_i)_{i \in N} \in \Theta^N$, $s_i^*(\theta_i)$ is dominant for each i in the resulting extensive-form game, and $f(\theta_1, \ldots, \theta_{|N|})$ coincides with the assignment that arises when players follow these strategies.

The main result of this section is a characterization of social choice functions that are implementable this way. As a start, we add a property, non-bossiness, that often appears in similar results for the rational domain. We extend the characterization to all implementable social choice functions further below. The social choice function f is *non-bossy* if the following holds: $f_j(\theta_i, \theta_{-i}) \neq f_j(\theta_i', \theta_{-i})$ implies $f_i(\theta_i, \theta_{-i}) \neq f_i(\theta_i', \theta_{-i})$, that is, a change in i's choice function impacts the assignment of another agent only if it also impacts i's assignment.

Let a *generalized serial dictatorship* be a tree $\mathcal{T}$ with an agent $i(n)$ and a finite set $A(n)$ of actions associated to each non-terminal node n, an assignment associated to each terminal node, and the following conditions:

(I) Each agent appears at most once along each path from the root to a terminal node;

(II) For each non-terminal node n, $A(n) \subseteq X^*$ and contains at least two elements.

(III) One cannot find an action $a(n') \in A(n) \setminus \{\emptyset\}$ on the path to n.

(IV) If action $a(n)$ appears on the path to a terminal node n', then $i(n)$ gets $a(n)$ at n'.

Classic serial dictatorship has each agent pick in turn according to a prespecified priority list from the entire set of items that haven't been selected by higher-priority agents. Similarly, agents in the above trees get to sequentially select at most once (by condition (I)) an element from a subset of X^* (the agent's choice determines her assignment by condition (IV)). Condition (III) ensures there is no path where a same item would be picked twice. But, contrary to serial dictatorship, an agent's turn and her opportunity set can vary with past agents' choices. Each such tree defines a social choice function, where the assignment at $(\theta_1, \ldots, \theta_{|n|})$ is simply the assignment at the terminal node reached when applying the choice functions $c_{\theta_1}, \ldots, c_{\theta_{|N|}}$ in the tree. Social choice functions of this type are called *generalized serial-dictatorship rules.*

Proposition 2. *A non-bossy social choice function is implementable in dominant strategies if, and only if, it is a generalized serial-dictatorship rule.*

Pápai (2000) and Pycia and Ünver (2017) provide a characterization of non-bossy and Pareto efficient social choice functions that are dominant-strategy implementable over the rational domain. Proposition 2 characterizes social choice functions that are dominant-strategy implementable over the unrestricted domain of all choice functions. We make a few comments as we compare these results.

Enlarging the domain of permissible behaviors makes implementation harder. First, agents have more deviations. Second, as argued in the previous section, the very existence of a dominant strategy becomes in some sense harder for given irrational choice-function profiles given that dominance against pure strategies (which underlies the very definition of strategyproofness) does not guarantee dominance against all beliefs in the absence of STP. We see that some rules identified by Pápai (2000) and Pycia and Ünver (2017), those involving non-trivial top-trading cycles, do not survive the stronger implementability requirement. But we also see, quite remarkably, that a substantial number of rules among those they identified are most robust in terms of their implementability. These rules are also particularly simple: agents sequentially make choices in opportunity sets instead of having to report complex messages (e.g., about what they would choose in different putative menus).

The Necessity of Dynamic Mechanisms A dynamic mechanism and its associated strategic form are viewed as equivalent when it comes to dominant-strategy implementation over the rational domain. By contrast, using dynamic mechanisms can be preferable over larger domains. Lessons from the previous section provide some intuition. Lesson 1 taught us that dominance is much more demanding than dominance against pure strategies in static games. But Lesson 2 brought some hope: dominance in a dynamic game can be easier than dominance in the associated strategic form by reducing the uncertainty agents face when making decisions. As should be clear by now, generalized serial dictatorships fully leverage this feature. The next result shows that most generalized serial dictatorship rule cannot be dominant-strategy implementable via a static mechanism (not the strategic-form associated to the dynamic game implementing it, nor *any* other static game). In fact, it provides a characterization of non-bossy social choice functions implementable this way.

Proposition 3. *A generalized serial dictatorship rule f is dominant-strategy implementable via a static mechanism if, and only if, $A(n) = A(n')$ for any two nodes n and n' such that $i(n) = i(n')$ (in the tree underlying the definition of f).*

Efficiency Contrary to typical results on the rational domain, our characterization does not rely on any form of efficiency. This is convenient since efficiency in the absence of preferences remains a much-debated topic. But this also raises interesting questions: which generalized serial-dictatorship rules are Pareto efficient on the rational domain, and what efficiency property (if any) do they entail on selected options for non-rational choice functions? An assignment α is (strongly) *Pareto efficient* for a rational choice-function profile if there does not exist another assignment α' such that all i for which $\alpha'_i \neq \alpha_i$ strictly prefer α'_i over α_i.

Proposition 4. *A generalized serial-dictatorship rule f is Pareto efficient on the rational domain if, and only if, the following properties hold in the tree underlying f:*

(i) The action set at the root is X^*;

(ii) If the path until a non-terminal node n contains at most $|N| - 2$ actions and action $a(n) = \emptyset$ is pursued, then the resulting node is also non-terminal and has the same action set ($A(n)$);

(ii) If the path until a non-terminal node n contains at most $|N| - 2$ actions, with at most $|X| - 2$ of them different from $\emptyset$, and action $a(n) \neq \emptyset$ is pursued, then the resulting node is also non-terminal and has the action set $A(n) \setminus \{a(n)\}$.

Efficiency thus bring us closer to classic serial dictatorship, though some more flexibility remains: who chooses at a node may depend on others' past choices instead of being predetermined via some priority ranking.[20] A couple of corollaries follow. First, combin-

[20] Pápai (2000) introduces such methods under the name of *sequential dictatorship* as examples of hierarchical exchange rules.

ing this proposition with the previous one, we conclude that no efficient generalized serial-dictatorship rule can be dominant-strategy implemented via a static mechanism. This illustrates once again the value of dynamic mechanisms beyond the rational domain. Second, we infer that Pareto efficiency over the rational domain guarantees efficiency à la Sugden (2004), Bernheim and Rangel (2009) and de Clippel (2014) over the unrestricted domain when it comes to generalized serial-dictatorship rules.[21]

Dropping Non-Bossiness Steps 1 and 2 in the proof of Proposition 2 characterize the set of social choice functions that are dominant-strategy implementable. Dropping non-bossiness allows one to consider more general trees. In a nutshell, an agent who selects an element in a subset of X^* may now have multiple actions associated to a same outcome for her. These otherwise redundant actions can impact the order of play for the remaining agents, as well as their opportunity sets. Similarly, an agent who had no say on her assignment in the generalized serial dictatorship could now have multiple actions that, though not impacting her lot, would once again impact the order of play for the remaining agents as well as their opportunity sets. For instance, the first mover might receive x for all choice function profiles, but her action choice determines the order of play among remaining agents over $X \setminus \{x\}$. The precise definition is available right after the proof of Step 2. While one may value the added flexibility, one issue with these additional mechanisms is that they'll have at least one agent with multiple dominant strategies. Thus, dropping non-bossiness does allow to implement additional social choice functions (though not tremendously more), but at the cost of relying on a dominant-strategy selection process that may seem rather ad-hoc. In the example just mentioned, one could require for instance ther first mover to pick a first order of other agents if her type is θ and another order if her type is θ', but she has no strict reason to comply.

References

Barberà, S., 2011. Strategyproof Social Choice. *Handbook of Social Choice and Welfare* **2**, 731-831.

[21] For our problem at hand, an assignment α is *efficient* given $(\theta_i)_{i\in N}$, according de Clippel (2014), if there exists a collection $(\mathcal{Z}_i)_{i\in N}$ of subsets of X^* such that (a) $\alpha_i = c_{\theta_i}(\mathcal{Z}_i)$, for each $i \in I$, and (b) for each assignment α', there exists at least one agent j for which $\mathcal{Z}_j \setminus \{\alpha_j\}$ contains α'_j. Conditions (i) to (iii) in Proposition 4 imply that this holds true by taking $\mathcal{Z}_j = A(n(j))$ where $n(j)$ is the node in the path followed at $(\theta_i)_{i\in N}$ such that $i(n(j)) = j$, if such a node exists, and $\mathcal{Z}_j = \{\emptyset\}$ if no such node exists. The reader is referred to Sections IV and V in de Clippel (2014) to see how this amounts to an instance of Sugden's (2004) "Opportunity Criterion" adapted in unit-demand problems, and a subset of Bernheim and Rangel's (2009) efficiency notion.

Bernhein, B. D., and A. Rangel, 2009. Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. *Quarterly Journal of Economics* **124**, 51-104.

de Clippel, G., 2014. Behavioral Implementation. *American Economic Review*, **104**, 2975-3002.

Gibbard, A., 1973. Manipulation of Voting Schemes: A General Result. *Econometrica* **41**, 587-601.

de Clippel, G., and K. Rozen, forthcoming. Bounded Rationality in Choice Theory: a Survey. *Journal of Economic Literature.*

Brandenburger, A., and E. Dekel, 1987. Rationalizability and Correlated Equilibria. *Econometrica* **55**, 1391-1402.

Dreyfuss, B., O. Heffetz, and M. Rabin, 2022. Expectations-Based Loss Aversion May Help Explain Seemingly Dominated Choices in Strategy-Proof Mechanisms. *American Economic Journal: Microeconomics* **14**, 515-555.

Dreyfuss, B., O. Glicksohn, O. Heffetz, and A. Romn, 2022. Deferred Acceptance with News Utility. *Mimeo.*

Esponda, I., and E. Vespa, 2021. Contingent Thinking and the Sure-Thing Principle: Revisiting Classic Anomalies in the Laboratory, *Mimeo.*

Hammond, P. J., 1979. Straightforward Individual Incentive Compatibility in Large Economies. *Review of Economic Studies*, **46**, 263-282.

Köszegi, B., and M. Rabin, 2009. Reference-Dependent Consumption Plans. *American Economic Review* **99**, 909-36.

Li, S., 2017. Obviously Strategy-Proof Mechanisms. *American Economic Review* **107**, 3257-87.

Mandal, P. and S. Roy, 2022. Obviously Strategy-Proof Implementation Of Assignment Rules: A New Characterization. *International Economic Review*, **63**, 261-290.

Manzini, P., and M. Mariotti, 2007. Sequentially Rationalizable Choice. *American Economic Review* **97**, 1824-1839.

Mas-Colell, A., M. D. Whinston, and J. R. Green, 1995. *Microeconomic Theory.* New York: Oxford University Press.

Pápai, S., 2000. Strategyproof assignment by hierarchical exchange. *Econometrica*, **68**, 1403-1433.

Pycia, M., and P. Troyan, forthcoming. A Theory of Simplicity in Games and Mechanism Design, *Econometrica.*

Pycia, M., and M. U. Ünver, 2017. Incentive Compatible Allocation and Exchange of Discrete Resources. *Theoretical Economics*, **12**, 287-329.

Satterthwaite, M. A., 1975. Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory* **10**, 187-217.

Savage, L., 1972. The Foundations of Statistics (2nd edition).

Sugden, R., 2004. The Opportunity Criterion: Consumer Sovereignty without the Assumption of Coherent Preferences. *American Economic Review*, **94**, 1014-1033.

Troyan, P., 2019. Obviously Strategy-Proof Implementation of Top Trading Cycles. *International Economic Review*, **60**, 1249-1261.

Appendix

Proof of Proposition 1

Proof. If c is not rational, then there exist a choice problem $T \subseteq O$ and $x \in T$ distinct from $c(T)$ such that $c(T) \neq c(T \setminus \{x\})$. Let Ω' be a nonempty strict subset of Ω, and consider the following acts:

	$\omega \in \Omega'$	$\omega \in \Omega \setminus \Omega'$
a	$c(T)$	$c(T \setminus \{x\})$
a'	$c(T \setminus \{c(T)\})$	$c(T \setminus \{x\})$
a''	$c(T \setminus \{x\})$	$c(T)$
a'''	x	$c(T)$
a_y	y	y

for each $y \in T \setminus \{x, c(T), c(T \setminus \{x\})\}$ (if any). Let A be the set of all acts appearing on the table. Then $A(\omega) = T$ for each $\omega \in \Omega'$, and $A(\omega) = T \setminus \{x\}$ for each $\omega \in \Omega \setminus \Omega'$. Let $p \in \Delta_{++}(\Omega)$ be such that $p(\Omega) = p(\Omega \setminus \Omega') = 1/2$. By STP, $c(L^p(A)) = \{\ell^p(a)\} = \{\frac{1}{2}c(T) \oplus \frac{1}{2}c(T \setminus \{x\})\}$. Let $\hat{A} = A \setminus \{a\}$. Then[22] $\hat{A}(\omega) = T \setminus \{c(T)\}$ for each $\omega \in \Omega'$, and $\hat{A}(\omega) = T \setminus \{x\}$ for each $\omega \in \Omega \setminus \Omega'$. By STP, $c(L^p(\hat{A})) = \{\ell^p(a')\} = \{\frac{1}{2}c(T \setminus \{c(T)\}) \oplus \frac{1}{2}c(T \setminus \{x\})\}$. A contradiction arises then from the fact that $L^p(A) = L^p(\hat{A})$ and $c(T) \neq c(T \setminus \{c(T)\})$. □

Proof of Proposition 2 The fact that generalized serial-dictatorship rules are implementable in dominant strategies is straightforward (simply by using the tree defining it as a game form to implement it). We proceed in two steps to prove necessity. Let f be a social choice function that is implementable in dominant strategies.

Step 1 *f is implementable in dominant strategies by a perfect-information game form $\mathcal{G}$ where agents play at most once along each path.*

Proof. Say that agent i *plays for the first time at her information set $\mathcal{I}$* if she does not have another information set $\mathcal{I}'$ such that $\mathcal{I}$ succeeds $\mathcal{I}'$.[23] Now modify the game form as follows. Each information set $\mathcal{I}$ at which an agent i plays for the first time in the original game form now becomes a singleton information set where she chooses an action in $S_i^{\mathcal{I}}$, her entire set of strategies starting at $\mathcal{I}$. All other information sets disappear. Transitions between new nodes, as well as final outcomes, are determined by implementing the new action choices as strategies in the original game form. Thus agents now pick strategies once and for all each

[22] The fact that $c(T \setminus \{x\}) \neq c(T)$ is important here.

[23] Of course, there can be multiple information sets at which an agent plays for the first time when i is not the first mover in the game as a whole. But, by perfect recall, the collections of information sets that succeed two such information sets are disjoint.

time they play for the first time, which may include specifying action choices at subsequent information sets of the original game form, and those strategies are observed by everyone.

To check that transitions are well defined, consider an agent i at a node ν of the new game form. By definition, this node corresponds to an information set $\mathcal{I}$ at which i plays for the first time in the original game form. When implementing i's chosen strategy and other players' strategies that led to ν, the original game form reaches either a terminal node, or a new information set $\mathcal{I}'$. Notice that $\mathcal{I}'$ cannot belong to i because her strategy at $\mathcal{I}$ specifies what to play at $\mathcal{I}'$, which succeeds $\mathcal{I}$. For the same reason, $\mathcal{I}'$ cannot belong to any agent that played before i on the path that led to ν in the new game form. This means that $\mathcal{I}'$ belongs to some agent j who plays there for the first time, which indeed corresponds to a new node in the new game. Thus transitions between nodes are indeed well defined, and agents play at most once along each path.

Consider now a choice function profile $(\theta_i)_{i\in N} \in \Theta^N$, and let $(s_i)_{i\in N}$ be the dominant strategy profile in the behavioral extensive-form game associated to the original game form. Implementing those strategies leads to a terminal node whose outcome is $f(\theta_1, \ldots, \theta_{|N|})$. These strategies translate naturally in the new game form: player i at a node ν corresponding to the information set $\mathcal{I}$ in the original game now picks the restriction of s_i to $\mathcal{I}$ and subsequent information sets, denoted $s_i^{\mathcal{I}}$. By construction, this strategy profile in the new game also leads to the final assignment $f(\theta_1, \ldots, \theta_{|N|})$.

It remains to check that these are dominant strategies in the modified game. Consider a node ν at which agent i picks an action in the new game form. We must check that[24] $c_{\theta_i}(\hat{\mathcal{O}}_i^\nu(\hat{\mu}_i^\nu)) = \{\hat{\ell}_i^\nu(\hat{s}_i^\nu, \hat{\mu}_i^\nu)\}$, for all probability distribution over the set of strategy profiles for others that are consistent with ν. Fixing notations, say ν corresponds to an information set $\mathcal{I}$ in the original game form. By definition, $\hat{s}_i^\nu = s_i^{\mathcal{I}}$. Notice that any strategy profile for others in the new game that is consistent with ν corresponds to a strategy profile for others in the original game that is consistent with $\mathcal{I}$ (the original game form can only accommodate more strategy profiles for others, as player i may be uncertain about what past movers will pick in the future if they happen to play again). Hence $\hat{\mu}_i^\nu$ admits a natural translation $\mu_i^{\mathcal{I}}$ in the original game form. We now have that $\hat{\ell}_i^\nu(\hat{s}_i^\nu, \hat{\mu}_i^\nu) = \ell_i^{\mathcal{I}}(s_i^{\mathcal{I}}, \mu_i^{\mathcal{I}})$, and $\hat{\mathcal{O}}_i^\nu(\hat{\mu}_i^\nu) = \mathcal{O}_i^{\mathcal{I}}(\mu_i^{\mathcal{I}})$. Given that $c_\theta(\mathcal{O}_i^{\mathcal{I}}(\mu_i^{\mathcal{I}})) = \{\ell_i^{\mathcal{I}}(s_i^{\mathcal{I}}, \mu_i^{\mathcal{I}})\}$, it must be that $c_{\theta_i}(\hat{\mathcal{O}}_i^\nu(\hat{\mu}_i^\nu)) = \{\hat{\ell}_i^\nu(\hat{s}_i^\nu, \hat{\mu}_i^\nu)\}$, as desired. □

Step 2 *Fix an information set $\mathcal{I}$ of $\mathcal{G}$ at which some agent i makes a move.*[25] *Then either i's choice at that node fully pins down her assigned element from X^* in the subgame, or i's choice at that node has no impact on the element she gets from X^* in the subgame.*

[24] Hats are used to denote variables associated to the new game form.

[25] $\mathcal{I}$ is a singleton since information is perfect in $\mathcal{G}$.

Proof. The result is proved by contradiction. Suppose to the contrary that there exist $s_i, s'_i, s''_i \in S^{\mathcal{I}}_i$, and $s_{-i}, s'_{-i}, s''_{-i} \in S^{\mathcal{I}}_{-i}$ such that (a) $o_i(s_i, s_{-i}) \neq o_i(s'_i, s_{-i})$ (i may impact her assigned element from X^* in the subgame) and (b) $o_i(s''_i, s'_{-i}) \neq o_i(s''_i, s''_{-i})$ (i's choice does not fully pin down her assigned element from X^* in the subgame).

We first prove this implies there exist $t_i, t'_i \in S^{\mathcal{I}}_i$ and $t_{-i}, t'_{-i} \in S^{\mathcal{I}}_{-i}$ such that (a') $o_i(t_i, t_{-i}) \neq o_i(t'_i, t_{-i})$ and (b') $o_i(t_i, t_{-i}) \neq o_i(t_i, t'_{-i})$. If there exists $r_{-i} \in S^{\mathcal{I}}_{-i}$ such that $o_i(s_i, s_{-i}) \neq o_i(s_i, r_{-i})$, then (a') and (b') hold with $t_i = s_i$, $t'_i = s'_i$, $t_{-i} = s_{-i}$ and $t'_{-i} = r_{-i}$. If there exists $r_{-i} \in S^{\mathcal{I}}_{-i}$ such that $o_i(s'_i, s_{-i}) \neq o_i(s'_i, r_{-i})$, then (a') and (b') hold with $t_i = s'_i$, $t'_i = s_i$, $t_{-i} = s_{-i}$ and $t'_{-i} = r_{-i}$. For the remaining cases we have that $o_i(s_i, s''_{-i}) = o_i(s_i, s_{-i}) \neq o_i(s'_i, s_{-i}) = o_i(s'_i, s''_{-i})$. Hence $o_i(s''_i, s''_{-i})$ cannot equal both $o_i(s_i, s''_{-i})$ and $o_i(s'_i, s''_{-i})$. Say that $o_i(s''_i, s''_{-i}) \neq o_i(s_i, s''_{-i})$ (a similar reasoning applies in the other case). Now (a') and (b') are satisfied with $t_i = s''_i$, $t'_i = s_i$, $t_{-i} = s''_{-i}$ and $t'_{-i} = s'_{-i}$.

Next, we show that, if two distinct strategy profiles $r_{-i}, r'_{-i} \in S^{\mathcal{I}}_{-i}$ are such that $\mathcal{O}^{\mathcal{I}}_i(r_{-i}) = \mathcal{O}^{\mathcal{I}}_i(r'_{-i})$, then i's assigned element from X^* does not depend on whether others play r_{-i} or r'_{-i}. Note that, for each $x \in \mathcal{O}^{\mathcal{I}}_i(r_{-i})$, there exists $r_i \in S^{\mathcal{I}}_i$ such that $o_i(r_i, r_{-i}) = o_i(r_i, r'_{-i}) = x$. Otherwise, i would select different strategies whether her belief is r_{-i} or r'_{-i} when her choice function selects x from $\mathcal{O}^{\mathcal{I}}_i(r_{-i})$, and the game would fail to have a dominant strategy. Furthermore, there cannot be $r_i \in S^{\mathcal{I}}_i$ such that $o_i(r_i, r_{-i}) \neq o_i(r_i, r'_{-i})$. Otherwise, the game would fail to have a dominant strategy for i at $\mathcal{I}$ given any choice function that selects the lottery $\frac{1}{2}o_i(s_i, r_{-i}) \oplus \frac{1}{2}o_i(s_i, r'_{-i})$ from $\mathcal{O}^{\mathcal{I}}_i(\frac{1}{2}r_{-i} \oplus \frac{1}{2}r'_{-i})$. Indeed, a dominant strategy would have to give i the same element from X^* in r_{-i} and r'_{-i} (since $\mathcal{O}^{\mathcal{I}}_i(r_{-i}) = \mathcal{O}^{\mathcal{I}}_i(r'_{-i})$), when using deterministic beliefs, and give a lottery with distinct elements from X^*, when using the belief $\frac{1}{2}r_{-i} \oplus \frac{1}{2}r'_{-i}$.

From the previous paragraph, we conclude that the $t_i, t'_i \in S^{\mathcal{I}}_i$ and $t_{-i}, t'_{-i} \in S^{\mathcal{I}}_{-i}$ identified earlier, namely which satisfy (a') and (b'), must be such that $\mathcal{O}^{\mathcal{I}}_i(t_{-i}) \neq \mathcal{O}^{\mathcal{I}}_i(t'_{-i})$. Consider now a choice function for i that selects $o_i(t_i, t_{-i})$ from $\mathcal{O}^{\mathcal{I}}_i(t_{-i})$ and $o_i(t_i, t'_{-i})$ from $\mathcal{O}^{\mathcal{I}}_i(t'_{-i})$ (which is possible because $\mathcal{O}^{\mathcal{I}}_i(t_{-i}) \neq \mathcal{O}^{\mathcal{I}}_i(t'_{-i})$). Assume furthermore that her choice function selects the lottery $\frac{3}{4}o_i(t'_i, t_{-i}) \oplus \frac{1}{4}o_i(t'_i, t'_{-i})$ from $\mathcal{O}^{\mathcal{I}}_i(\frac{3}{4}t_{-i} \oplus \frac{1}{4}t'_{-i})$ (which is possible because the latter set is distinct from both $\mathcal{O}^{\mathcal{I}}_i(t_{-i})$ and $\mathcal{O}^{\mathcal{I}}_i(t'_{-i})$ given (b')). In that case, a dominant strategy for i at $\mathcal{I}$ gives $o_i(t_i, t_{-i})$ when her belief is t_{-i}, $o_i(t_i, t'_{-i})$ when her belief is t'_{-i}, and $\frac{3}{4}o_i(t'_i, t_{-i}) \oplus \frac{1}{4}o_i(t'_i, t'_{-i})$ when her belief is $\frac{3}{4}t_{-i} \oplus \frac{1}{4}t'_{-i}$. But no strategy can do this given (a'). □

Steps 1 and 2 reveal that f is dominant-strategy implementable using a tree with an agent $i(n)$ and a finite set $A(n)$ of actions associated to each non-terminal node n, an assignment associated to each terminal node, and the following properties:

(I) For each n, there is a map $g_{|N|}$ from $A(n)$ to X^*. Let $Im(g_{|N|})$ denote the image of $A(n)$;

(II) Each agent appears at most once along each path from the root to a terminal node;

(III) One cannot find an action $a(m)$ on the path to n such that $g_m(a(m)) \in Im(g_{|N|}) \setminus \{\emptyset\}$;

(IV) If action $a(n)$ appears on the path to a terminal node n', then $i(n)$ gets $g_{|N|}(a(n))$ at n'.

Dominant strategies are easily determined in these trees, for all choice-function profiles: a strategy for agent i is dominant given θ_i if, and only if, the action $a(n)$ selected at each node n such that $i(n) = i$ is such that $c_{\theta_i}(Im(g_{|N|})) = g_{|N|}(a(n))$.[26]

Step 3 *If f is, in addition, non-bossy, then f is a generalized serial-dictatorship rule.*

Proof. Fix one of the above game form to implement f in dominant strategies. Let $(s_i^* : \Theta \to S_i)_{i \in N}$ be a profile of dominant strategies implementing f. Let n be a node in the tree such that (i) $A(n)$ is strictly larger than $Im(g_{|N|})$, and (ii) there is no subsequent node m such that $A(m)$ is strictly larger than $Im(g_m)$. We now show that f is dominant strategy implementable in a modified game form where the only change occurs is that each element of X^* selectable at n is now associated to a unique action. The result then follows by repeatedly adjusting the game form this way (a backward-induction argument).

To do this, we trim twice the action set at n. First, we eliminate any action $a(n)$ that is never played at n, that is, for which one cannot find $\theta_{i(n)}$ such that $s^*_{i(n)}(\theta_{i(n)}) = a(n)$. Let $A'(n) \subseteq A(n)$ be the set of surviving actions. Notice that $g_{|N|}(A'(n)) = g_{|N|}(A(n))$. Otherwise, for $x \in g_{|N|}(A(n)) \setminus g_{|N|}(A'(n))$, $i(n)$ would select an action in $A(n) \setminus A'(n)$, contradicting the definition of eliminated actions, when her choice function is rational and ranks x at the top

Next, for each $x \in Im(g_{|N|})$, fix an element $a_x^* \in A'(n)$ such that $g_{|N|}(a_x^*) = x$, and consider the modified set $A^*(n) = \{a_x^* | x \in Im(g_{|N|})\} \subseteq A'(n) \subseteq A(n)$. The function $g_{|N|}$ remains unchanged (though now defined over a smaller domain), and the new tree is obtained simply by trimming subgames associated to actions in $A(n) \setminus A^*(n)$. Clearly, $g_{|N|}(A^*(n)) = g_{|N|}(A'(n))$, and hence $g_{|N|}(A^*(n)) = g_{|N|}(A(n))$. Dominant strategies are modified by selecting a_x^* whenever $i(n)$ picked $a(n)$ in the original game such that $g_{|N|}(a(n)) = x$.

Modified strategies are clearly dominant since opportunity sets, as subsets of X^*, remain unchanged at all nodes surviving the trimming. It remains to show that they continue

[26] This provides a characterization of social choice functions that are dominant-strategy implementable (without non-bossiness).

to implement f. For this, fix a choice function profile $(\theta_i)_{i\in N}$. Let $x = g_{|N|}(s_i^*(\theta_{i(n)})) = f_i(\theta_1, \ldots, \theta_{|N|})$ be the element of X^* that $i(n)$ selects in the original game form, and hence receives under f, when of type $\theta_{i(n)}$. By definition of A^*, there exists $\theta'_{i(n)}$ such that $s_i^*(\theta'_{i(n)}) = a_x^*$. Hence the final assignment when playing the new dominant strategies at $(\theta_i)_{i\in N}$ in the modified game gives rise to the same assignment as when playing the original dominant strategies at $(\theta'_{i(n)}, \theta_{-i(n)})$ in the original game, or $f(\theta'_{i(n)}, \theta_{-i(n)})$. But $f_i(\theta'_{i(n)}, \theta_{-i(n)}) = f_i(\theta_{i(n)}, \theta_{-i(n)}) = x$, since $g_{|N|}(a_x^*) = x$, and hence $f_i(\theta'_{i(n)}, \theta_{-i(n)}) = f_i(\theta_{i(n)}, \theta_{-i(n)})$, since f is non-bossy. Thus the modified game form, along with the modified strategies, also implement f in dominant strategies, as desired. □

Proof of Proposition 3

Proof. (*Sufficiency*) For each $j \in N$, let A_j be the set of actions available at all n such that $i(n) = j$. Consider then the static mechanism where agents select, independently of each others, an element of their action set, and the outcome associated to a strategy profile is the outcome obtained by implementing those strategies in the underlying tree (a simplification of the associated strategic form). Clearly, selecting $c_{\theta_j}(A_j)$ is dominant for each player j of type θ_j, and this static mechanism implements f in dominant strategies.

(*Necessity*) We prove the contraposition. Suppose i is in charge at n and n' such that $a \in A(n) \setminus A(n')$ (a similar argument applies if one can only find an action in $A(n') \setminus A(n)$). Suppose also that $A(n')$ contains a' and a'' (action sets contain at least two elements by condition (II) in the definition of the tree underlying the definition of f). For any path in the tree, we can construct a choice-function profile generating this path. Indeed, each action corresponds to an element of X^* that the agent in charge can select, and the unrestricted domain contains a choice function that would pick that element from the opportunity set arising when varying those actions. Fix now θ_i such that i picks a from $A(n)$ and a' from $A(n')$, θ'_i such that i picks a from $A(n)$ and a'' from $A(n')$, θ_{-i} such that n is reached, and θ'_{-i} such that n' is reached.

Suppose, for an argument by contradiction, that f is dominant-strategy implementable via a static mechanism. Let M_j be the set of strategies available to each player j, and let $(m_j : \Theta \to M_j)_{j\in N}$ be a profile of dominant strategies implementing f. If agent i picks the message $m_i(\theta_i)$ in the static mechanism, then she gets a (resp., a') if j picks $m_j(\theta_j)$ (resp., $m_j(\theta'_j)$) for each $j \neq i$. If agent i picks the message $m_i(\theta'_i)$ in the static mechanism, then she gets a (resp., a'') if j picks $m_j(\theta_j)$ (resp., $m_j(\theta'_j)$) for each $j \neq i$. Now suppose that agent i has a choice function θ''_i that selects a from $A(n)$, a' from $A(n')$, and $\frac{1}{2}a \oplus \frac{1}{2}a''$ from $\{\frac{1}{2}g_i(w_i, (m_j(\theta_j))_{j\neq i}) \oplus \frac{1}{2}g_i(w_i, (m_j(\theta'_j))_{j\neq i}) | w_i \in M_i\}$, where g is the outcome function in the static mechanism implementing f (the selected lottery does indeed belong on that set, simply taking $w_i = m_i(\theta'_i))$. The dominant strategy $m_i(\theta''_i)$ would have to be such that $g_i(m_i(\theta''_i), (m_j(\theta_j))_{j\neq i}) = a$ (dominance for i's belief that other agents play according to $(m_j(\theta_j))_{j\neq i}$), $g_i(m_i(\theta''_i), (m_j(\theta'_j))_{j\neq i}) = a'$ (dominance for i's belief that other agents play according to $(m_j(\theta'_j))_{j\neq i}$), and $\frac{1}{2}g_i(m_i(\theta''_i), (m_j(\theta_j))_{j\neq i}) \oplus \frac{1}{2}g_i(m_i(\theta''_i), m_j(\theta'_j))_{j\neq i}) = \frac{1}{2}a \oplus \frac{1}{2}a''$ (dominance for i's belief that other agents play according to $(\frac{1}{2}m_j(\theta_j) \oplus \frac{1}{2}m_j(\theta'_j))_{j\neq i})$. This is impossible given that $a' \neq a''$, which contradicts the fact that the static mechanism admits a dominant strategy. □

Proof of Proposition 4

Proof. (*Necessity*) Suppose n^* is the root of the tree, but that there exists $y \in X^* \setminus A(n^*)$. Consider then a rational choice-function profile where all agents other that $i(n^*)$ rank $\emptyset$ at the top, while agent $i(n^*)$ ranks y at the top. Pareto efficiency implies that agents other than $i(n^*)$ get $\emptyset$, while implementability implies that $i(n^*)$ gets her top choice in $A(n^*)$. But this is Pareto dominated by the assignment that give y to $i(n^*)$ (and $\emptyset$ to all other agents). This establishes (i).

We now prove (ii) and (iii) by induction on the length of the path at n. Suppose thus that (i), (ii) and (iii) hold at all previous nodes along the path that led to n, in addition to the fact that the path until n contains at most $|N| - 2$ actions. Suppose first that $a(n) = \emptyset$ is pursued, but that the action set at the resulting node n' is different from $A(n)$. By the induction hypothesis, it means that $A(n')$ is a strict subset of $A(n)$ (since a chosen item is eliminated forever after along each path of the tree). Say $y \in A(n) \setminus A(n')$. Consider then a rational choice-function profile where $i(n')$ ranks y at the top, all her predecessors top rank the action they follow on that path, and all remaining agents rank $\emptyset$ at the top. Implementability implies that $i(n')$ gets her top choice in $A(n')$ and her predecessors get the action followed on the path until n'. Pareto efficiency implies that all remaining agents get $\emptyset$. But this is Pareto dominated by modifying the assignment to give y to $i(n')$ instead. This establishes (ii). A very similar argument applies to establish (iii).

(*Sufficiency*) Suppose the tree underlying the definition of f satisfies conditions (i) to (iii), but that f selects a Pareto inefficient assignment at some rational choice-function profile $(\theta_i)_{i\in N}$. Let α be a Pareto improving assignment, and let n be the first node at which $\alpha_{i(n)}$ is different from the action chosen at $i(n)$ when following the tree to implement f at $(\theta_i)_{i\in N}$. By properties (i) to (iii), $\alpha_{i(n)} \in A(n)$, since $A(n)$ is obtained by eliminating options other than $\alpha_{i(n)}$ (the assignment α cannot assign $\alpha_{i(n)}$ to both $i(n)$ and one of her predecessors) from X^*. But having $i(n)$ chooses $f_{i(n)}(\theta_1, \ldots, \theta_{|N|})$ from $A(n)$ contradicts the fact that $i(n)$ strictly prefers $\alpha_{i(n)}$ over $f_{i(n)}(\theta_1, \ldots, \theta_{|N|})$. □