Ex-post Behavioral Implementation

Mehmet Barlo* Nuh Aygün Dalkıran†

February 20, 2026

Abstract

We provide necessary as well as sufficient conditions for ex-post behavioral implementation under incomplete information. Ex-post consistency, the central condition we identify as necessary and almost sufficient for ex-post behavioral implementation, is a natural extension of ex-post incentive compatibility and ex-post monotonicity to behavioral domains. As an application, we adapt the notion of ex-post incentive efficiency in Holmström and Myerson (1983) to behavioral environments and show that the resulting social choice correspondence is fully ex-post behavioral implementable under mild conditions.

Keywords: Behavioral Implementation, Incomplete Information, Ex-Post Implementation, Ex-post Behavioral Incentive Efficiency.

JEL Classification: C72, D60, D79, D82, D90

*Corresponding author: Faculty of Arts and Social Sciences, Sabancı University, Tuzla, İstanbul, 34956, Türkiye; +90 (216) 483 9284; barlo@sabanciuniv.edu. ORCID: 0000-0001-6871-5078

†Department of Economics, Bilkent University, Çankaya, Ankara, 06800, Türkiye; +90 (312) 290 2006; dalkiran@bilkent.edu.tr. ORCID: 0000-0002-0586-0355

Table of Contents

1 Introduction

Individuals are not perfect decision-makers. They often have difficulty processing information and making rational choices, as documented by behavioral sciences. Behavioral economics uses the resulting insights and aims to help governments and other organizations design policies and institutions and guide people's behavior more effectively. In this context, we focus on how a planner can achieve her goals under incomplete information when individuals are not necessarily making rational choices, and their interim behavior can be captured by ex-post considerations. Specifically, we study full ex-post behavioral implementation without requiring individuals' ex-post and interim choices to satisfy the weak axiom of revealed preferences (WARP).

We employ *ex-post behavioral equilibrium* (EBE) to derive necessary as well as sufficient conditions for *ex-post behavioral implementation* of social choice rules. This notion requires every optimal state-contingent alternative be sustained as an EBE, and there be no EBE leading to an outcome not aligned with the social choice rule.

In incomplete information environments, each mechanism induces an incomplete information game where, given a strategy profile, each individual's message generates an interim act that maps each type profile of other individuals to alternatives. As a result, we obtain a general setup with incomplete information that allows for a wide variety of behavioral biases. In this setup, individuals make their message choices in the mechanism at the interim stage (after observing their own private information).

The notion of EBE requires a strategy profile in the mechanism be such that individuals' plans of actions are measurable with respect to their private information and result in (behavioral) Nash equilibrium play at every state.[1]

In behavioral environments of incomplete information, combining an individual's optimal ex-post choices across states does not necessarily induce optimality in the interim stage for that individual. We present a condition, *Property STP**, in the spirit of Savage's sure-thing principle as well as Property STP of de Clippel (2023) in order to relate individuals' ex-post choices on alternatives to their interim choices on acts. It demands that

[1]Full ex-post behavioral implementation is not the same as behavioral (Nash) implementation on every complete information type space: Each individual's strategy is measurable with respect to her type and cannot vary with others' type profiles. This, therefore, results in a requirement akin to the addition of incentive compatibility constraints. The associated necessary conditions are not nested even in the rational domain (Bergemann & Morris, 2008). Under rationality, our necessary conditions coincide with those of Bergemann and Morris (2008). Besides, Bergemann and Morris (2005) shows that even though partial ex-post implementation is equivalent to (interim) Bayesian implementation for all possible type spaces in some environments, this equivalence does not hold in the case of full implementation.

an act be chosen from a set of acts whenever for any state, the realization of this act (an alternative) is ex-post chosen at that state from the set of alternatives sustained by acts in that set of acts. Consequently, we obtain the practicality and tractability of the ex-post approach featured in the rational domain. Under Property STP*, every EBE is a behavioral interim equilibrium (BIE) (Saran, 2011; Barlo & Dalkıran, 2023a). Hence, ex-post behavioral implementation shares some desirable properties of its counterpart in the rational domain with Savage-Bayesian probabilistic sophistication where Property STP* comes for free (Bergemann & Morris, 2008).[2] In particular, we obtain the *ex-post no regret property*: no individual has any incentive to go back to the interim stage and find out others' private information.

We obtain *ex-post consistency* as a necessary condition for ex-post behavioral implementation, Theorem 1, and present a sufficiency result, Theorem 2, combining this condition with a behavioral version of the economic environment assumption. This central condition, necessary and almost sufficient for ex-post behavioral implementation, is a natural extension of ex-post incentive compatibility and ex-post monotonicity to behavioral domains: We show that ex-post consistency implies *quasi-ex-post incentive compatibility* (Proposition 1) and *ex-post behavioral monotonicity* (Proposition 2). However, these two concepts together do not imply ex-post consistency in the behavioral domain (Example 2). In contrast, in the rational domain with utility representation, our quasi-ex-post incentive compatibility and ex-post behavioral monotonicity are equivalent to the ex-post incentive compatibility and ex-post monotonicity of Bergemann and Morris (2008) (Proposition 3). Moreover, ex-post consistency is equivalent to ex-post monotonicity coupled with ex-post incentive compatibility in the rational domain with utility representation (Proposition 4).

To illustrate an application of our results, we study the implementability of *ex-post behavioral efficiency* and *ex-post behavioral incentive efficiency*. Under rationality, the former coincides with ex-post Pareto efficiency, while the latter refines the notion of ex-post incentive Pareto efficiency introduced by Holmström and Myerson (1983). We first show that, in behavioral environments, the classic tension between efficiency and incentives persists, implying that ex-post behavioral efficiency is not ex-post behaviorally implementable. Our necessity results indicate that any implementable efficiency notion

[2]While Property STP* instigates appealing aspects for EBE, it comes with a severe warning in environments with individuals' ex-post choices failing WARP. In such environments, de Clippel (2023) exhorts us to be wary of the use of EBE because Property STP* and the failure of WARP for ex-post choices may generate a contradiction. In Appendix C, we demonstrate situations in which such a contradiction may appear in our setup. We note that interim choices may satisfy Property STP* but fail WARP even if the associated ex-post choices obey WARP, as in the minimax-regret setting (Proposition 7).

must, at a minimum, incorporate a behavioral analogue of ex-post incentive compatibility. Imposing this restriction leads to ex-post behavioral incentive efficiency. We then show that the corresponding opportunity sets form an ex-post consistent profile and, by our sufficiency result (Proposition 5), conclude that ex-post behavioral incentive efficiency is ex-post behaviorally implementable.

Ex-post behavioral implementation features some robustness properties with respect to individuals' beliefs: Given a set of permissible type spaces as in Bergemann and Morris (2011), if a social choice rule is ex-post behavioral implementable at a permissible type space, then it is ex-post behavioral implementable at every other permissible type space. In this sense, ex-post behavioral implementability is independent of individuals' beliefs. However, ex-post behavioral implementability does not dismiss the occurrence of bad BIE at a permissible type space regardless of whether or not Property STP* holds. This is why one has to beware of the warning of Bergemann and Morris (2008) about the "naive adoption of ex-post [behavioral] equilibrium as a 'robust' solution concept."[3]

Our paper is most closely related to Bergemann and Morris (2008) and Barlo and Dalkıran (2023a). The former analyzes ex-post implementation under incomplete information in the rational domain, while the latter studies interim implementation under incomplete information in behavioral environments, without imposing any ex-post requirements. Our results on ex-post behavioral implementation extend these analyses by introducing explicit ex-post considerations into behavioral domains.

In the rational domain, Jackson (1991) develops Bayesian implementation, extending the seminal Nash implementation framework of Maskin (1999) to settings with incomplete information. Finally, the equilibrium concept of BIE is introduced by Saran (2011), who studies partial behavioral implementation under incomplete information in environments with menu-dependent preferences.

The necessary and almost-sufficient conditions we identify for ex-post behavioral implementation are reminiscent of the consistency condition introduced by de Clippel (2014), which provides necessary and sufficient conditions for behavioral implementation under complete information. A related literature studies behavioral implementation in complete-information environments, including Hurwicz (1986), Eliaz (2002), Barlo and Dalkiran

[3]We analyze the robust behavioral implementation problem in Barlo and Dalkıran (2026). Bergemann and Morris (2009, 2011) analyze robust full implementation in the rational domain both using direct and indirect mechanisms. There is a large literature on robust mechanism design/implementation in the rational domain. Such studies include but are not limited to Bergemann and Morris (2005), Penta (2015), Bergemann and Morris (2017), Ollár and Penta (2017), de Clippel et al. (2019), Kunimoto et al. (2025), Chen et al. (2021), Chen et al. (2022), Jain and Lombardi (2022), Jain et al. (2025), Chen et al. (2023), Jain et al. (2023), Kunimoto et al. (2023), Xiong (2023), Kunimoto and Saran (2024).

(2009), Korpela (2012), Hayashi et al. (2023), and Hagiwara (2025). Eliaz (2002) analyzes full implementation when some individuals may be "faulty" and fail to act optimally. An earlier paper of ours, Barlo and Dalkiran (2009), studies implementation under ε-Nash equilibrium, allowing individuals to settle for outcomes close to, but not necessarily achieving, their best responses. Korpela (2012) shows that when individual choice behavior violates rationality axioms, independence of irrelevant alternatives plays a key role in recovering necessary and sufficient conditions analogous to those in Moore and Repullo (1990). More recently, Hayashi et al. (2023) studies behavioral strong implementation under complete information, while Hagiwara (2025) analyzes behavioral subgame-perfect implementation.[4]

The organization of the paper is as follows: Section 2 presents the preliminaries; Section 3, our necessity and sufficiency results for ex-post behavioral implementation. In Section 4, we analyze ex-post behavioral (incentive) efficiency. Section 5 discusses implications of the sure thing principle. Section 6 concludes.

2 Preliminaries

Consider a set of individuals $N = \{1, \ldots, n\}$ and a non-empty set of alternatives X where 2^X stands for the set of all subsets of X and $\mathcal{X}$ stands for those that are non-empty. Let T denote the set of all relevant type profiles of the individuals. We assume that there is incomplete information among the individuals regarding the true state of the world and that their information is exclusive. Thus, T has a product structure, i.e., $T := \times_{i \in N} T_i$ where $t_i \in T_i$ denotes the type of individual $i \in N$. When individual i observes her type t_i at the interim stage, her choice (payoff) type θ_i and her assessments (beliefs) about others' types are determined. We let Θ_i be the set of all possible choice-types of individual i and assume that the realized choice-type of individual i is determined by the *surjective* function $\vartheta_i : T_i \to \Theta_i$.[5] Let $\Theta := \times_{i \in N} \Theta_i$ be the set of all choice-states. Individual i's *ex-post choice correspondence at type profile* t is described by $c_i^{\vartheta(t)} : \mathcal{X} \to 2^X$, with $c_i^{\vartheta(t)}(S) \subseteq S$ for all $S \in \mathcal{X}$ where $\vartheta(t) := (\vartheta_1(t_1), \ldots, \vartheta_n(t_n))$. The ex-post environment denoted by $\mathcal{E}^{\text{ep}}$ specifies the *type space* $(T, \vartheta) = (T_i, \vartheta_i)_{i \in N}$ as detailed above. We assume that $\mathcal{E}^{\text{ep}}$ is

[4] Additional related contributions include Kucuksenel (2012), Ohashi (2012), Korpela and Lombardi (2019), Barlo and Dalkıran (2022, 2023b), and Rubbini (2023).

[5] To analyze the robust implementation problem in the rational domain, Bergemann and Morris (2011) considers type spaces where an individual's type specifies not only her payoff type but also her probabilistic beliefs about the types of the other individuals. This is why they define a type space as a triple $\mathcal{T} = (T_i, \hat{\theta}_i, \hat{\pi}_i)_{i \in N}$ where $\hat{\theta}_i : T_i \to \Theta_i$ specifies the payoff type of individual i, and $\hat{\pi}_i : T_i \to \Delta(T_{-i})$ specifies her beliefs about the types of the other individuals where $\hat{\pi}_i(t_{-i} \mid t_i) \in [0, 1]$ denotes the belief of individual i of type t_i about the other individuals' type profile being t_{-i}.

common knowledge among the individuals. We impose *no restrictions* on ex-post choices such as WARP.[6] In particular, we allow individuals' ex-post choices to be empty valued unless explicitly stated otherwise.

We restrict attention to social choice functions (SCFs) determined solely by choice-states as in Bergemann and Morris (2011). An SCF is $h : \Theta \to X$ mapping Θ into X, and we denote the set of all SCFs by $H := \{h \mid h : \Theta \to X\}$. Because a planner may consider many socially optimal SCFs simultaneously, we consider social choice sets (SCSs). An SCS F is a non-empty set of SCFs, i.e., $F \subset H$ and $F \neq \emptyset$; an SCF $f \in F$ specifies a socially optimal alternative—as evaluated by the planner—for each choice-state.[7]

A *mechanism* is given by $\mu = (M, g)$ where M_i denotes individual i's non-empty set of *messages* with $M = \times_{i \in N} M_i$, and $g : M \to X$ describes the *outcome function* identifying the alternative corresponding to each message profile. A mechanism induces an incomplete information game form in our environment. Given the type space (T, ϑ) specified by the ex-post environment $\mathcal{E}^{\text{ep}}$, a *strategy* of individual i in mechanism μ, $\sigma_i : T_i \to M_i$, specifies a message for each type of i.

Individual i's opportunity set of alternatives under mechanism μ for a given message profile of other individuals $m_{-i} \in M_{-i}$ equals $O_i^{\mu}(m_{-i}) := \{g(m_i, m_{-i}) \in X \mid m_i \in M_i\}$.

Any mechanism that implements an SCS under incomplete information must consider individuals' private information. However, individuals may be deceitful. Given an ex-post environment, we denote a *deception* by individual i as $\alpha_i : T_i \to T_i$. Intuitively, $\alpha_i(t_i)$ can be thought of as individual i's reported type. Consequently, $\alpha(t) := (\alpha_1(t_1), \alpha_2(t_2), \ldots, \alpha_n(t_n))$ designates a profile of possibly deceitful reported types while α^{id} denotes the *truthtelling profile*, i.e., $\alpha_i^{\text{id}}(t_i) = t_i$ for all $i \in N$ and all $t_i \in T_i$. We denote individual i's set of all possible deceptions by Λ_i and let $\Lambda := \times_{i \in N} \Lambda_i$, $\Lambda_{-i} := \times_{j \neq i} \Lambda_j$, and $\alpha_{-i}(t_{-i}) := (\alpha_j(t_j))_{j \neq i}$.

3 Ex-post Behavioral Implementation

First, we present the definition of the concept of ex-post equilibrium in a given ex-post environment:

[6] Sen (1971) shows that a choice correspondence satisfies WARP (and is represented by a complete and transitive preference relation) if and only if it satisfies independence of irrelevant alternatives (IIA) and an expansion consistency axiom (known as Sen's β). Letting $\mathcal{Z}$ be the set of all non-empty subsets of alternatives, we say that the choice correspondence $c : \mathcal{Z} \to \mathcal{Z}$ satisfies (i) the IIA if $x \in S \cap c(T)$ for some $S, T \in \mathcal{Z}$ with $S \subset T$ implies $x \in c(S)$; (ii) Sen's β if $x, y \in S \subset T$ for some $S, T \in \mathcal{Z}$, and $x, y \in c(S)$ implies $x \in c(T)$ if and only if $y \in c(T)$.

[7] We note that it is customary to denote a social choice rule as an SCS rather than a social choice correspondence under incomplete information. We refer the interested reader to Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1987), Jackson (1991), and Bergemann and Morris (2008).

Definition 1. *Given an ex-post environment and hence a type space (T, ϑ), a strategy profile $\sigma^* = (\sigma_1^*, \ldots, \sigma_n^*)$ is an* **ex-post behavioral equilibrium** (EBE) *of mechanism μ if for every type profile $t \in T$, we have $g(\sigma^*(t)) \in c_i^{\vartheta(t)}(O_i^\mu(\sigma_{-i}^*(t_{-i})))$ for all $i \in N$.*

In words, an EBE requires the outcomes generated by the mechanism be a (behavioral) Nash equilibrium at every type profile, while individuals' strategies have to be measurable with respect to only their own types.[8]

Definition 2. *Given an ex-post environment, an SCS $F \in \mathcal{F}$ is* **ex-post behavioral implementable** *if there exists a mechanism μ such that*

(i) for every $f \in F$, there exists an EBE σ^* of μ such that $f \circ \vartheta = g \circ \sigma^*$, and

(ii) for every EBE σ^* of μ, there exists $f \in F$ such that $g \circ \sigma^* = f \circ \vartheta$.

The following exemplifies the notion of ex-post behavioral implementation via a direct mechanism in an ex-post environment with rational interdependent preferences where the type profiles and choice-states are in one-to-one correspondence. The main purpose of this example is to display how our necessary condition, *ex-post consistency*, deals with deceptions that result in non-trivial complications even in this relatively straightforward setting under incomplete information.[9,10]

Example 1. Let $N = \{1, 2\}$, $X = \{x, y, z\}$, $T_i = \{t_i, t_i'\}$ and $\vartheta_i(t_i) = \theta_i$ and $\vartheta_i(t_i') = \theta_i'$ for both $i = 1, 2$. Table 1 details the rational interdependent ex-post preferences (rankings).

$R_{1,(\theta_1,\theta_2)}$	$R_{2,(\theta_1,\theta_2)}$	$R_{1,(\theta_1',\theta_2)}$	$R_{2,(\theta_1',\theta_2)}$	$R_{1,(\theta_1,\theta_2')}$	$R_{2,(\theta_1,\theta_2')}$	$R_{1,(\theta_1',\theta_2')}$	$R_{2,(\theta_1',\theta_2')}$
x	x	z	z	z	z	y	y
z	z	x	x	y	y	z	z
y	y	y	y	x	x	x	x

Table 1: Ex-post preferences

Using the construction that we present in Section 4, we see that in this example, the only ex-post behavioral incentive efficient SCF f is given by $f(\theta_1, \theta_2) = x$, $f(\theta_1, \theta_2') = f(\theta_1', \theta_2) = z$, and $f(\theta_1', \theta_2') = y$, which we denote by $\langle xzzy \rangle$.

[8] A message profile $m^* \in M$ is a (behavioral) Nash equilibrium of μ at t if $g(m^*) \in \cap_{i \in N} c_i^{\vartheta(t)}(O_i^\mu(m_{-i}^*))$.

[9] In Section 5, we extend this example to a minimax-regret framework. In that setting, interim choices violate WARP even though ex-post choices remain rational. More generally, allowing ex-post choices to violate WARP while imposing a sure-thing-principle–type restriction linking interim and ex-post behavior may lead to an impasse, as illustrated by de Clippel (2023) and discussed in detail in Appendix C. Finally, motivated by the appeal of direct mechanisms, we provide an analysis of ex-post behavioral implementation via direct mechanisms in Appendix B.

[10] In Appendix A, we present another example where ex-post choices violate WARP, an indirect mechanism ex-post behavioral implements a given SCS, and the revelation principle fails for EBE.

An intuitive interpretation of our example involves a firm's headquarters (HQ, the planner) consisting of two subdivisions (individuals) $i = 1, 2$. The subdivisions are located in two separate countries. The HQ needs to extract the state pertaining to the economic outlook of country i from each division separately. Each subdivision's type is in one-to-one correspondence with its country's economic state (choice-type). Naturally, each subdivision is informed about its type but not that of the other. Each country's economic state (and hence its subdivision's choice-type) is either 'good' or 'bad' where θ_i denotes the former and θ_i' the latter. There are three possible firm-wide policies (alternatives) the HQ is to adopt: *expansion*, *contraction*, and *prudence*, denoted by x, y, and z, respectively.

As subdivisions are parts of the same organization, their state-contingent (rational) ex-post preferences equal one another at each choice-state, resulting in a common-value-like setting with interdependent preferences. In particular, if both countries' choice-types are good, each subdivision ranks *expansion* strictly over *prudence* and *prudence* strictly over *contraction*; if both countries' choice-types are bad, each strictly prefers *contraction* to *prudence* and *prudence* to *expansion*. On the other hand, if the choice-type of one country is good and the other is bad, then each strictly top-ranks *prudence* while (i) if country 2 (involving a bigger market when compared to that of country 1) is in a good state, and country 1's is bad, then each strictly ranks *expansion* over *contraction*; (ii) otherwise, each strictly prefers *contraction* to *prudence*.

Intuitively, the HQ's state-contingent goal, $F = \{f\}$, involves *expansion* if both countries' states are good, *contraction* if both countries' states are bad, and *prudence* for every other possible situation.

The informational advantages of ex-post implementation are evident in this example: We do not need to specify subdivisions' and the HQ's beliefs about others' types. Yet, we see that the (direct) mechanism in Table 2 ex-post behavioral implements $F = \{f\}$.

		Individual 2	
		t_2	t_2'
Individual 1	t_1	x	z
	t_1'	z	y

Table 2: The mechanism that ex-post behavioral implements $F = \{f\}$.

We wish to highlight that individuals' ability to use deceptions results in complications: Even when type profiles and choice-states are in one-to-one correspondence and we use direct mechanisms, 16 SCFs emerge during the play of the mechanisms via deceptions. To see such a complication, consider the strategy profile given by deception $\alpha^{(11)}$ in Table 3 where both individuals $i = 1, 2$ claim to be of type t_i' regardless of their true types. Then, the resulting SCF is $\langle yyyy \rangle$ as $f(\vartheta(\alpha^{(11)}(t))) = f(\theta_1', \theta_2') = y$ for all $t \in T$.

Table 3 lists all deceptions and resulting SCFs, and establishes that the truthtelling strategy profile, α^{id}, is the only EBE of the direct mechanism. Under α^{id}, every type

	$\alpha_1(t_1)$	$\alpha_1(t'_1)$	$\alpha_2(t_2)$	$\alpha_2(t'_2)$	$f \circ \vartheta \circ \alpha$	Informant state	Informant (Deviator)	Conforming outcome	Deviation outcome
α^{id}	t_1	t'_1	t_2	t'_2	$\langle xzzy \rangle$	–	–	–	–
$\alpha^{(2)}$	t_1	t'_1	t_2	t_2	$\langle xzxz \rangle$	(t'_1, t'_2)	2	z	y
$\alpha^{(3)}$	t_1	t'_1	t'_2	t'_2	$\langle zyzy \rangle$	(t_1, t_2)	2	z	x
$\alpha^{(4)}$	t_1	t'_1	t'_2	t_2	$\langle zyxz \rangle$	(t'_1, t'_2)	2	z	y
$\alpha^{(5)}$	t_1	t_1	t_2	t'_2	$\langle xxzz \rangle$	(t'_1, t_2)	1	x	z
$\alpha^{(6)}$	t_1	t_1	t_2	t_2	$\langle xxxx \rangle$	(t'_1, t_2)	1	x	z
$\alpha^{(7)}$	t_1	t_1	t'_2	t'_2	$\langle zzzz \rangle$	(t_1, t_2)	2	z	x
$\alpha^{(8)}$	t_1	t_1	t'_2	t_2	$\langle zzxx \rangle$	(t_1, t_2)	2	z	x
$\alpha^{(9)}$	t'_1	t'_1	t_2	t'_2	$\langle zzyy \rangle$	(t_1, t_2)	1	z	x
$\alpha^{(10)}$	t'_1	t'_1	t_2	t_2	$\langle zzzz \rangle$	(t'_1, t'_2)	2	z	y
$\alpha^{(11)}$	t'_1	t'_1	t'_2	t'_2	$\langle yyyy \rangle$	(t_1, t_2)	1	y	z
$\alpha^{(12)}$	t'_1	t'_1	t'_2	t_2	$\langle yyzz \rangle$	(t_1, t_2)	2	y	z
$\alpha^{(13)}$	t'_1	t_1	t_2	t'_2	$\langle zxyz \rangle$	(t_1, t_2)	1	z	x
$\alpha^{(14)}$	t'_1	t_1	t_2	t_2	$\langle zxzx \rangle$	(t_1, t_2)	1	z	x
$\alpha^{(15)}$	t'_1	t_1	t'_2	t'_2	$\langle yzyz \rangle$	(t_1, t_2)	1	y	z
$\alpha^{(16)}$	t'_1	t_1	t'_2	t_2	$\langle yzzx \rangle$	(t_1, t_2)	2	y	z

Table 3: Deception profiles and corresponding informants

of every individual reveals their type truthfully and obtains the commonly top-ranked alternative at every state, establishing that α^{id} is an EBE of μ. To see that there is no other EBE strategy profile, consider $\alpha^{(7)}$ as an example, where both types of individual 1 claim to be of type t_1 while both types of individual 2 claim to be of type t'_2. Thus, SCF $\langle zzzz \rangle$ emerges as z is the resulting alternative at every type profile. Consequently, by conforming to $\alpha^{(7)}$, individual 2 of type t_2 (who is to claim to be of type t'_2) obtains the alternative z at type profile (t_1, t_2). However, if individual 2 of type t_2 deviates to truthtelling at the interim stage (and claim to be of type t_2), she attains alternative x at (t_1, t_2), top-ranked at that profile. So, individual 2 of type t_2 has a 'strictly profitable' deviation opportunity under $\alpha^{(7)}$ and hence can serve as an informant for this deception.[11]

3.1 Necessity

A necessary condition for ex-post behavioral implementation is *ex-post consistency*:

Definition 3. *Given an ex-post environment, a profile of sets of alternatives given by* $\mathbb{S} := (S_i(f, \theta_{-i}))_{i \in N,\ f \in F,\ \theta_{-i} \in \Theta_{-i}}$ *is* **ex-post consistent with the SCS** F *if for every SCF* $f \in F$*,*

(*i*) *for all* $i \in N$ *and all* $t \in T$, $f(\vartheta(t)) \in c_i^{\vartheta(t)}(S_i(f, \vartheta_{-i}(t_{-i})))$*, and*

(*ii*) *for any deception profile* α *and* $\tilde{f} \notin F$ *with* $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$*, there are* $t^* \in T$ *and* $i^* \in N$ *such that* $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*}))))$.

[11] We discuss how these observations relate to ex-post consistency, our necessary condition, following its definition below.

A profile of sets of alternatives $\mathbb{S}$ is ex-post consistent with an SCS F if the following hold: (i) Given any $i \in N$ and any $f \in F$ and any $t \in T$, it must be that i's ex-post choice at $\vartheta(t)$ from the corresponding choice set $S_i(f, \vartheta_{-i}(t_{-i}))$ contains $f(\vartheta(t))$; (ii) given any $f \in F$, whenever there is a deception profile α that leads to an outcome not compatible with the SCS, there exist a type profile t^* and an informant i^* such that $f(\vartheta(\alpha(t^*)))$ is not in the ex-post choice of i^* at $\vartheta(t^*)$ from $S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*})))$.

To illustrate ex-post consistency through Example 1, we note that the profile of sets $\mathbb{S} = (S_i(f, \theta_{-i}))_{i \in N,\ f \in F,\ \theta_{-i} \in \Theta_{-i}}$ with $S_i(f, \theta_j) = \{x, z\}$ and $S_i(f, \theta'_j) = \{y, z\}$ for $i, j = 1, 2$ with $i \neq j$ is ex-post consistent with the SCS $F = \{f\}$. This follows from (i) $f(\vartheta(t_i), \vartheta(t_j)) = x \in c_i^{\vartheta(t_i, t_j)}(\{x, z\})$, $f(\vartheta(t'_i), \vartheta(t_j)) = z \in c_i^{\vartheta(t'_i, t_j)}(\{x, z\})$, $f(\vartheta(t_i), \vartheta(t'_j)) = z \in c_i^{\vartheta(t_i, t'_j)}(\{y, z\})$, $f(\vartheta(t'_i), \vartheta(t'_j)) = y \in c_i^{\vartheta(t'_i, t'_j)}(\{y, z\})$ for both $i = 1, 2$ while (ii) for each deception profile $\alpha^{(k)}$ for $k = 2, \ldots, 16$ leading to an SCF not aligned with SCS F, the informant individual i^* and the informant state t^* that satisfies (ii) of ex-post consistency are as given in Table 3.

If mechanism μ ex-post behavioral implements a given SCS F in an ex-post environment, then for any SCF $f \in F$, there is an EBE σ^f of μ such that $f \circ \vartheta = g \circ \sigma^f$. Thus, for all $t \in T$, $g(\sigma^f(t)) = f(\vartheta(t)) \in \cap_{i \in N} c_i^{\vartheta(t)}(O_i^\mu(\sigma^f_{-i}(t_{-i})))$. Defining $\mathbb{S}$ by $S_i(f, \vartheta_{-i}(t_{-i})) := O_i^\mu(\sigma^f_{-i}(t_{-i}))$ for all $i \in N$, $f \in F$, and $t \in T$ implies (i) of ex-post consistency of $\mathbb{S}$ with F. Further, if a deception profile α is such that for some $\tilde{f} \notin F$ we have $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$, then $\sigma^f \circ \alpha$ cannot be an EBE of μ; because otherwise, by (ii) of ex-post implementability, $\tilde{f} \circ \vartheta = g \circ \sigma^f \circ \alpha$, which implies $\tilde{f} \in F$, a contradiction. So, there is a type profile t^* and an informant i^* whose ex-post choice at $\vartheta(t^*)$ from $O_{i^*}^\mu(\sigma^f_{-i^*}(\alpha_{-i^*}(t^*_{-i^*})))$ (which equals $S_{i^*}(f, \vartheta_{i^*}(\alpha_{-i^*}(t^*_{-i^*})))$) does not include $f(\vartheta(\alpha(t^*))) = \tilde{f}(\vartheta(t^*))$. This delivers (ii) of ex-post consistency of $\mathbb{S}$ with F. This discussion proves the following necessity result for ex-post behavioral implementation:

Theorem 1. *Given an ex-post environment, if an SCS F is ex-post behavioral implementable, then there is a profile of sets of alternatives ex-post consistent with F.*

To establish that our ex-post necessity result extends the analysis of Bergemann and Morris (2008) to behavioral domains, we show that our necessary condition, ex-post consistency, implies *analogs* of theirs: *quasi-ex-post incentive compatibility* and *ex-post behavioral monotonicity*. Notably, however, these two conditions together do not imply ex-post consistency in the behavioral domain, as we demonstrate in Example 2. Then, we display that in the rational domain with utility representation, our quasi-ex-post incentive compatibility and ex-post behavioral monotonicity are equivalent to the ex-post incentive compatibility and ex-post monotonicity of Bergemann and Morris (2008). Further,

we prove that ex-post consistency is equivalent to ex-post monotonicity combined with ex-post incentive compatibility in the rational domain with utility representation.

Given an ex-post environment, an SCS F is ***quasi-ex-post incentive compatible*** if for every SCF $f \in F$, type profile $t \in T$, and individual $i \in N$, there is a set of alternatives $S \in \mathcal{X}$ such that $f(\vartheta(t)) \in c_i^{\vartheta(t)}(S)$ and $f(\Theta_i, \vartheta_{-i}(t_{-i})) \subseteq S$ where $f(\Theta_i, \vartheta_{-i}(t_{-i})) := \{f(\theta_i', \vartheta_{-i}(t_{-i})) \in X \mid \theta_i' \in \Theta_i\}$.

Proposition 1. *Given an ex-post environment, if there exists a profile of sets of alternatives ex-post consistent with an SCS F, then F is quasi-ex-post incentive compatible.*

To see the arguments needed to establish this result, let $\mathbb{S} := (S_i(f, \theta_{-i}))_{i \in N,\ f \in F,\ \theta_{-i} \in \Theta_{-i}}$ be a profile of sets of alternatives ex-post consistent with F. Given any $t \in T$ and $i \in N$, set $S := S_i(f, \vartheta_{-i}(t_{-i}))$. By (i) of ex-post consistency, $f(\vartheta(t)) \in c_i^{\vartheta(t)}(S)$, establishing the first condition of quasi-ex-post incentive compatibility. Since $f(\theta_i', \vartheta_{-i}(t_{-i})) \in c_i^{(\theta_i', \vartheta_{-i}(t_{-i}))}(S_i(f, \vartheta_{-i}(t_{-i})))$ for each $\theta_i' \in \Theta_i$ by (i) of ex-post consistency, $f(\theta_i', \vartheta_{-i}(t_{-i})) \in S$ for each $\theta_i' \in \Theta_i$, establishing $f(\Theta_i, \vartheta_{-i}(t_{-i})) \subseteq S$.

Given an ex-post environment, an SCS F is ***ex-post behavioral monotonic*** if for all SCF $f \in F$, deception profile α, and $\tilde{f} \notin F$ with $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$, there is a type profile $t^* \in T$, an individual $i^* \in N$, and a set of alternatives $S^* \in \mathcal{X}$ such that

(i) $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(S^*)$, and

(ii) $f(\theta_{i^*}', \vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*}))) \in c_{i^*}^{(\theta_{i^*}', \vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*})))}(S^*)$ for all $\theta_{i^*}' \in \Theta_{i^*}$.

Proposition 2. *Given an ex-post environment, if there exists a profile of sets of alternatives ex-post consistent with an SCS F, then F is ex-post behavioral monotonic.*

Proposition 2 directly follows from the existence of a profile of sets of alternatives that are ex-post consistent with the given SCS F: Given a profile of sets of alternatives $\mathbb{S} := (S_i(f, \theta_{-i}))_{i \in N,\ f \in F,\ \theta_{-i} \in \Theta_{-i}}$ ex-post consistent with F, let $S^* := S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*})))$. Then, (i) of ex-post behavioral monotonicity follows from (ii) of ex-post consistency while (ii) of ex-post behavioral monotonicity follows from (i) of ex-post consistency since $\vartheta_i : T_i \to \Theta_i$ is surjective for all $i \in N$.

However, quasi-ex-post incentive compatibility and ex-post behavioral monotonicity do *not* imply ex-post consistency in behavioral domains as we demonstrate in Example 2.

Example 2. Let $N = \{1, 2\}$, $X = \{x, y, z\}$, $T_1 = \{t_1\}$, $T_2 = \{t_2', t_2''\}$, $\Theta_1 = \{\theta_1\}$, $\Theta_2 = \{\theta_2', \theta_2''\}$ with $\vartheta((t_1, t_2')) = (\theta_1, \theta_2')$ and $\vartheta((t_1, t_2'')) = (\theta_1, \theta_2'')$. Suppose SCS F contains a single SCF, i.e., $F = \{f\}$ with $f(\theta_1, \theta_2') = x$ and $f(\theta_1, \theta_2'') = y$. For our purposes, it suffices to specify the ex-post choices of individual 1 and 2 as in Table 4.

	$c_1^{(\theta_1,\theta_2')}$	$c_2^{(\theta_1,\theta_2')}$	$c_1^{(\theta_1,\theta_2'')}$	$c_2^{(\theta_1,\theta_2'')}$
$\{x,y,z\}$	$\{x\}$	$\{x\}$	$\{y\}$	$\{z\}$
$\{x,y\}$	$\{x\}$	$\{y\}$	$\{y\}$	$\{y\}$

Table 4: Ex-post choices of Individuals 1 and 2.

SCS F is quasi-ex-post incentive compatible: Observe that the ex-post choices of individual 1 are fully aligned with SCF f. Thus, for both states, letting $S = \{x, y, z\}$, satisfies the required condition for individual 1, i.e., $x \in c_1^{(\theta_1,\theta_2')}(\{x,y,z\})$ and $y \in c_1^{(\theta_1,\theta_2'')}(\{x,y,z\})$. On the other hand, the only set that individual 2 chooses x at (θ_1, θ_2') that includes $f(\theta_1, \Theta_2) = \{x, y\}$ is $\{x, y, z\}$ whereas the only set that individual 2 chooses y at (θ_1, θ_2'') that includes $f(\theta_1, \Theta_2) = \{x, y\}$ is $\{x, y\}$, i.e., $x \in c_2^{(\theta_1,\theta_2')}(\{x,y,z\})$ and $y \in c_2^{(\theta_1,\theta_2'')}(\{x,y\})$.

Next, we show that SCS F satisfies ex-post behavioral monotonicity. Because there is only one type of individual 1, there are only four possible deceptions. Table 5 presents the informant state t^*, the informant individual i^*, and the informant set S^* satisfying (i) and (ii) of ex-post behavioral monotonicity for each of the deceptions that lead to an outcome not aligned with SCS $F = \{f\}$.

	$\alpha_1(t_1)$	$\alpha_2(t_2')$	$\alpha_2(t_2'')$	$f \circ \vartheta \circ \alpha$	Informant state t^*	Informant Individual i^*	Informant Set S^*
α^{id}	t_1	t_2'	t_2''	$\langle xy \rangle$	–	–	–
$\alpha^{(2)}$	t_1	t_2'	t_2'	$\langle xx \rangle$	(t_1, t_2'')	1	$\{x,y\}$
$\alpha^{(3)}$	t_1	t_2''	t_2''	$\langle yy \rangle$	(t_1, t_2')	1	$\{x,y\}$
$\alpha^{(4)}$	t_1	t_2''	t_2'	$\langle yx \rangle$	(t_1, t_2')	1	$\{x,y\}$

Table 5: Deception profiles and corresponding informants

To see that no profile of sets is ex-post consistent with SCS $F = \{f\}$, suppose for contradiction such a profile exists. Then, condition (i) of ex-post consistency for $i = 2$ and $t = (t_1, t_2')$ implies $S_2(f, \theta_1)$—the set associated with individual 2, SCF f, and θ_1—must satisfy $x \in c_2^{(\theta_1,\theta_2')}(S_2(f, \theta_1))$. This means the only candidate for $S_2(f, \theta_1)$ is $\{x, y, z\}$. However, condition (i) of ex-post consistency cannot then hold for $i = 2$ and $t = (t_1, t_2'')$, as $y \notin c_2^{(\theta_1,\theta_2'')}(\{x,y,z\})$. Consequently, such a $S_2(f, \theta_1)$ does not exist, proving our claim.

To analyze ex-post implementation in the rational domain when ex-post choices can be rationalized by utility functions, we denote the utility of individual $i \in N$ from alternative $x \in X$ at payoff-state $\theta \in \Theta$ by $u_i(x, \theta)$, and let $c_i^\theta(S) := \{y \in S : u_i(y, \theta) \geq u_i(x, \theta) \text{ for all } x \in S\}$ for any $S \in \mathcal{X}$.[12] Then, the necessary conditions of Bergemann and Morris (2008) are as follows: An SCS F is ***ex-post incentive compatible*** if for every

[12]It is well-known that WARP is equivalent to utility representation with finitely many alternatives, while some additional regularity assumptions may be needed with infinitely many alternatives.

$f \in F$ and every $t \in T$, $u_i(f(\vartheta(t)), \vartheta(t)) \geq u_i(f(\theta_i', \vartheta_{-i}(t_{-i})), \vartheta(t))$ for all $i \in N$ and all $\theta_i' \in \Theta_i$. Meanwhile, an SCS F is ***ex-post monotonic*** if for any $f \in F$, any α, and any $\tilde{f} \notin F$ with $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$, there exist $i \in N$, $t \in T$, and $y \in X$ such that

(*i*) $u_i(y, \vartheta(t)) > u_i(f(\vartheta(\alpha(t))), \vartheta(t))$, and

(*ii*) $u_i(f(\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))), (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))) \geq u_i(y, (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))))$ for all $\theta_i' \in \Theta_i$.

We now establish that in the rational domain with utility representation, the necessary conditions of Bergemann and Morris (2008) are equivalent to our ex-post behavioral monotonicity coupled with quasi-ex-post incentive compatibility.

Proposition 3. *In the rational domain with utility representation, ex-post behavioral monotonicity coupled with quasi-ex-post incentive compatibility is equivalent to ex-post monotonicity coupled with ex-post incentive compatibility.*

Proof of Proposition 3. The result follows from Claims 1 and 2.

Claim 1. *In the rational domain with utility representation, quasi-ex-post incentive compatibility is equivalent to ex-post incentive compatibility.*

Proof. Suppose that individuals' ex-post choices are represented by utility functions. If F is quasi-ex-post incentive compatible, then for all $f \in F$, all $t \in T$, and all $i \in N$, there exists $S \in \mathcal{X}$ such that $f(\Theta_i, \vartheta_{-i}(t_{-i})) \subset S$ and $f(\vartheta(t)) \in c_i^{\vartheta(t)}(S)$. Hence, the definition of $c^{\vartheta(t)}$ under utility representation implies $u_i(f(\vartheta(t)), \vartheta(t)) \geq u_i(f(\theta_i', \vartheta_{-i}(t_{-i})), \vartheta(t))$ for all $\theta_i' \in \Theta_i$, i.e., F is ex-post incentive compatible. Conversely, if F is ex-post incentive compatible, then for all $f \in F$, all $t \in T$, and all $i \in N$, $u_i(f(\vartheta(t)), \vartheta(t)) \geq u_i(f(\theta_i', \vartheta_{-i}(t_{-i})), \vartheta(t))$ for all $\theta_i' \in \Theta_i$. Letting $S = f(\Theta_i, \vartheta_{-i}(t_{-i}))$ delivers the result. ■

Claim 2. *In the rational domain with utility representation, the following hold:*

(*i*) *if an SCS F is ex-post behavioral monotonic, then it is ex-post monotonic, and*

(*ii*) *if an SCS F is ex-post monotonic and ex-post incentive compatible, then it is ex-post behavioral monotonic.*

Proof. Suppose that individuals' ex-post choices are represented by utility functions.

For (*i*), suppose for any $f \in F$, any α, and any $\tilde{f} \notin F$ with $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$, there exist $i \in N$, $t \in T$, and $S \in \mathcal{X}$ such that $f(\vartheta(\alpha(t))) \notin c_i^{\vartheta(t)}(S)$ while $f(\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))) \in c_i^{(\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i})))}(S)$ for all $\theta_i' \in \Theta_i$. Let $y \in c_i^{\vartheta(t)}(S)$. Then, by the definition of ex-post choices $c_i^{\vartheta(t)}$ under utility representation, we have $u_i(y, \vartheta(t)) > u_i(f(\vartheta(\alpha(t)), \vartheta(t))$ and

$u_i(f(\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))), (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))) \geq u_i(y, (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i})))$ for all $\theta_i' \in \Theta_i$. This establishes that F is ex-post monotonic.

For (ii), for any $f \in F$, any α, and any $\tilde{f} \notin F$ with $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$, there exist $i \in N$, $t \in T$, and $y \in X$ such that $u_i(y, \vartheta(t)) > u_i(f(\vartheta(\alpha(t))), \vartheta(t))$ and $u_i(f(\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))), (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))) \geq u_i(y, (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i})))$ for all $\theta_i' \in \Theta_i$. Let $S = f(\Theta_i, \vartheta_{-i}(\alpha_{-i}(t_{-i}))) \cup \{y\}$. Note that $f(\vartheta(\alpha(t))) \in S$ and (by the definition of ex-post choices) $f(\vartheta(\alpha(t))) \notin c_i^{\vartheta(t)}(S)$. Since F is ex-post incentive compatible by hypothesis, for all $\theta_i' \in \Theta_i$ we have that $u_i(f(\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))), (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))) \geq u_i(f(\tilde{\theta}_i, \vartheta_{-i}(\alpha_{-i}(t_{-i}))), (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i})))$ for all $\tilde{\theta}_i \in \Theta_i$ and $u_i(f(\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))), (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))) \geq u_i(y, (\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i})))$. Hence, $f(\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i}))) \in c_i^{(\theta_i', \vartheta_{-i}(\alpha_{-i}(t_{-i})))}(S)$ for all $\theta_i' \in \Theta_i$. Therefore, F is ex-post behavioral monotonic. ■ ■

Finally, we show that ex-post monotonicity coupled with ex-post incentive compatibility (alternatively, ex-post behavioral monotonicity and quasi-ex-post incentive compatibility) coincides with ex-post consistency under rationality with utility representation.

Proposition 4. *In the rational domain with utility representation, there is a profile of sets ex-post consistent with SCS F if and only if F is ex-post incentive compatible and satisfies ex-post monotonicity.*

Proof. ($\Rightarrow$) This direction follows from Propositions 1, 2, and 3.

($\Leftarrow$) Suppose F is ex-post incentive compatible and satisfies ex-post monotonicity. Let $R_{i,\theta}$ and $P_{i,\theta}$ represent the weak and strict preferences of i that rationalizes i's ex-post choices. Consider the profile of sets $\mathbb{S} := (S_i(f, \theta_{-i}))_{i \in N,\ f \in F,\ \theta_{-i} \in \Theta_{-i}}$ where for all $i \in N$, all $f \in F$, and all $\theta \in \Theta$, $S_i(f, \theta_{-i}) := \bigcap_{\theta_i' \in \Theta_i} LCS_i(f(\theta_i', \theta_{-i}), (\theta_i', \theta_{-i}))$ where $LCS_i(f(\theta_i', \theta_{-i}), (\theta_i', \theta_{-i})) := \{y \in X \mid f(\theta_i', \theta_{-i})\ R_{i,(\theta_i', \theta_{-i})}\ y\}$ for all $i \in N$, all $f \in F$, and all $\theta \in \Theta$. By construction, for all $i \in N$, all $f \in F$, and all $\theta \in \Theta$, every alternative in $S_i(f, \vartheta_{-i}(t_{-i}))$ is in the lower-contour set of individual i for $f(\vartheta(t))$ at $\vartheta(t)$. Hence, for all $i \in N$ and all $t \in T$, $f(\vartheta(t)) \in c_i^{\vartheta(t)}(S_i(f, \vartheta_{-i}(t_{-i})))$, i.e., (i) of ex-post consistency holds.

Let α be a deception such that $\tilde{f} \notin F$ with $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$. Then, by ex-post monotonicity, there exist $i^* \in N$, $t^* \in T$, and $y \in X$ such that $y\ P_{i^*, \vartheta(t^*)}\ f(\vartheta(\alpha(t^*))$, and $f(\theta_{i^*}', \vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*})))\ R_{i^*, (\theta_{i^*}', \vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*})))}\ y$ for all $\theta_{i^*}' \in \Theta_{i^*}^*$. Therefore, we have $y \in LCS_{i^*}(f(\theta_{i^*}', \vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*}^*))), (\theta_{i^*}', \vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*}))))$ for all $\theta_{i^*}' \in \Theta_{i^*}$. Then, $y \in S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*}^*)))$ by construction. But, because $y\ P_{i^*, \vartheta(t^*)}\ f(\vartheta(\alpha(t^*))$ this implies $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*}^*))))$. Therefore, there are $t^* \in T$ and $i^* \in N$ such that $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*}^*))))$, i.e., (ii) of ex-post consistency holds. ■

3.2 Sufficiency

We need the following to establish our sufficiency result:

Definition 4. *The* **ex-post choice incompatibility** *holds in an ex-post environment if for all $t \in T$, all $x \in X$, all $\bar{j} \in N$, there is $i^* \in N \setminus \{\bar{j}\}$ such that $x \notin c_{i^*}^{\vartheta(t)}(X)$.*

This condition implies some level of disagreement among individuals regarding their ex-post choices at every type profile.

Ex-post choice incompatibility coupled with ex-post consistency is sufficient for ex-post behavioral implementation:

Theorem 2. *Suppose that the ex-post environment is such that $n \geq 3$ and the ex-post choice incompatibility holds. Then, if there is a profile of sets of alternatives ex-post consistent with SCS F, then F is ex-post behavioral implementable.*

Before we proceed to the proof, we would like to note that under rationality, our ex-post choice incompatibility is equivalent to the economic environment assumption of Bergemann and Morris (2008).[13] Consequently, thanks to Propositions 1, 2, 3, and 4, Theorem 2 amounts to a behavioral analog of Bergemann and Morris's Theorem 2.

Proof of Theorem 2. Suppose that $n \geq 3$ and the ex-post choice incompatibility holds in the given ex-post environment. Consider SCS F with which the profile of sets of alternatives $\mathbb{S} := (S_i(f, \theta_{-i}))_{i \in N,\ f \in F,\ \theta_{-i} \in \Theta_{-i}}$ is ex-post consistent.

We use the following mechanism $\mu = (M, g)$: For each $i \in N$, her set of messages is $M_i = \big(F \cup \{\emptyset\}\big) \times \Theta_i \times X \times N$, while a generic message is denoted by $m_i = (m_i^1, \theta_i, x_i, k_i)$, and the outcome function $g : M \to X$ is as specified in Table 6.

Rule 1 :	$g(m) = f(\theta)$		if $m_i = (f, \theta_i, \cdot, \cdot)$ for all $i \in N$,
Rule 2 :	$g(m) = \begin{cases} x_j & \text{if } x_j \in S_j(f, \theta_{-j}), \\ f(\tilde{\theta}_j, \theta_{-j}) & \text{otherwise.} \end{cases}$		if $m_i = (f, \theta_i, \cdot, \cdot)$ for all $i \in N \setminus \{j\}$ and $m_j = (m_j^1, \tilde{\theta}_j, x_j, \cdot)$ with $m_j^1 \neq f$,
Rule 3 :	$g(m) = x_{j^*}$	where $j^* = \sum_i k_i \pmod n$	otherwise.

Table 6: The outcome function of the mechanism.

Claim 3. *For any $f \in F$, there exists an EBE, σ^f, of $\mu = (M, g)$ with $f \circ \vartheta = g \circ \sigma^f$.*

[13] Ex-post choice incompatibility is equivalent to the behavioral version of Bergemann and Morris (2008)'s economic environment assumption: For each $t \in T$ and each $x \in X$, there exist $i, j \in N$ with $i \neq j$ such that $x \notin c_i^{\vartheta(t)}(X)$ and $x \notin c_j^{\vartheta(t)}(X)$; i.e., at any type profile, any alternative can be ex-post chosen from the set of all alternatives by at most $n - 2$ individuals.

Proof. Let $\sigma_i^f(t_i) = (f, \vartheta_i(t_i), x, 1)$ for each $i \in N$ and for some arbitrary $x \in X$. By Rule 1, we have $g(\sigma^f(t)) = f(\vartheta(t))$ for each $t \in T$, i.e., $f \circ \vartheta = g \circ \sigma^f$. Observe that for any unilateral deviation by individual i from σ^f, either Rule 1 or Rule 2 applies, i.e., Rule 3 is not attainable by any unilateral deviation from σ^f. By construction, $O_i^\mu(\sigma_{-i}^f(t_{-i})) = S_i(f, \vartheta_{-i}(t_{-i}))$ for each $t \in T$, $i \in N$. Since, by (i) of ex-post consistency, $f(\vartheta(t)) \in c_i^{\vartheta(t)}(S_i(f, \vartheta_{-i}(t_{-i})))$ for each $i \in N$, we have for each $t \in T$, $g(\sigma^f(\vartheta(t))) \in c_i^{\vartheta(t)}(O_i^\mu(\sigma_{-i}^f(t_{-i})))$ for all $i \in N$, i.e., σ^f is an EBE of μ such that $f \circ \vartheta = g \circ \sigma^f$. ■

Consider now any EBE σ^* of μ denoted as $\sigma_i^*(t_i) = (m_i^1(t_i), \alpha_i(t_i), x_i(t_i), k_i(t_i))$ for each $i \in N$. That is, $m_i^1(t_i)$ denotes either a flag—designated as $\emptyset$—or the SCF proposed by i when her type is t_i; $\alpha_i(t_i)$, the reported type of i when her type is t_i; $x_i(t_i)$, the alternative proposed by i when her type is t_i; and $k_i(t_i)$, the number proposed by i when her type is t_i.

Claim 4. *Under any EBE σ^* of μ, Rule 1 must apply at each $t \in T$, and hence there is unique $f \in F$ with $m_i^1(t_i) = f$ for all $i \in N$ and all $t_i \in T_i$.*

Proof. Suppose, for contradiction, that either Rule 2 or Rule 3 applies at some $\tilde{t} \in \Theta$ under σ^*. If Rule 2 applies at $\tilde{t}$, by construction, we have $O_j^\mu(\sigma_{-j}^*(\tilde{t}_{-j})) = S_j(f, \vartheta_{-j}(\alpha_{-j}(\tilde{t}_{-j})))$ for the odd-man-out $j \in N$ and $O_i^\mu(\sigma_{-i}^*(\tilde{t}_{-i})) = X$ for all $i \neq j$, i.e., for all the other $n-1$ individuals.[14] On the other hand, if Rule 3 applies at $\tilde{t}$, we have, by construction, $O_i^\mu(\sigma_{-i}^*(\tilde{t}_{-i})) = X$ for all $i \in N$. In this case, simply let $j = 1$. Therefore, under both Rule 2 and Rule 3, at least $n-1$ individuals have the opportunity set X. Since σ^* is an EBE of μ, it follows that $g(\sigma^*(\tilde{t})) \in c_i^{\vartheta(\tilde{t})}(X)$ for all $i \neq j$. Consequently, the desired contradiction emerges due to ex-post choice incompatibility because there is no $i^* \neq j$ with $g(\sigma^*(\tilde{t})) \notin c_{i^*}^{\vartheta(\tilde{t})}(X)$.

As Rule 1 applies at every $t \in T$ under any EBE σ^* of μ, due to the product structure of T, there must be unique $f \in F$ with $f_i(t_i) = f$ for all $i \in N$ and all $t_i \in T_i$. Hence, by Rule 1, $g(\sigma^*(t)) = f(\vartheta(\alpha(t)))$ for all $t \in T$. ■

Claim 5. *For any EBE σ^* of μ, $\tilde{f}$ such that $g \circ \sigma^* = \tilde{f} \circ \vartheta$ is in F.*

Proof. Rule 1 applies at each $t \in T$, and each $i \in N$ reports the type $\alpha_i(t_i) \in T_i$ as the second entry of their messages at $t \in T$ under σ^*. Consequently, $\tilde{f}$ described in

[14] We note that an individual other than the odd-man-out may be unable to secure all alternatives under Rule 2 unless the flag $\emptyset$ is available—i.e., a transition from Rule 2 to Rule 3 may not be feasible without the use of the flag. To illustrate this, consider the case where $N = \{1, 2, 3\}$, $F = \{f, f'\}$, and the message profile is given by $m_1 = f$, $m_2 = f$, and $m_3 = f'$. In this setting, individuals 1 and 2 cannot trigger a transition to Rule 3 through unilateral deviations in the absence of the flag.

the claim is such that $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$ where f is the unanimously announced SCF under σ^*. Then, by construction, at each $t \in T$, $O_i^{\mu}(\sigma^*_{-i}(t_{-i})) = S_i(f, \vartheta_{-i}(\alpha_{-i}(t_{-i})))$ for all $i \in N$. By (ii) of ex-post consistency, if $\tilde{f} \notin F$, then there are $t^* \in T$ and $i^* \in N$ such that $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(S_{i^*}(f, \vartheta_{-i}(\alpha_{-i^*}(t_{-i^*}))))$. But this implies $g(\sigma^*(t^*)) \notin c_{i^*}^{\vartheta(t^*)}(O_{i^*}^{\mu}(\sigma^*_{-i^*}(t^*_{-i^*})))$, a contradiction to σ^* being an EBE of μ. ■ ■

4 Ex-post Behavioral Efficiency

In this section, we introduce a behavioral counterpart of ex-post incentive Pareto efficiency of Holmström and Myerson (1983). Our construction parallels de Clippel (2014), introducing the following efficiency notion in behavioral domains of complete information: An alternative is *behaviorally efficient* at a state of the world if each individual has an implicit opportunity set from which she chooses this alternative at that state, and each alternative is in at least one of these implicit opportunity sets. Extending this efficiency notion to incomplete information environments, we define *ex-post behavioral efficiency* by demanding such SCFs result in behaviorally efficient alternatives at every type profile:

Definition 5. *Given an ex-post environment, an SCF $f : \Theta \to X$ is **ex-post behavioral efficient** if there is a profile of sets of alternatives $(Y_{i,\theta})_{i \in N, \theta \in \Theta}$ such that*

(i) for all $i \in N$ and all $t \in T$, $f(\vartheta(t)) \in c_i^{\vartheta(t)}(Y_{i,\vartheta(t)})$, and

(ii) for all $t \in T$, $\cup_{i \in N} Y_{i,\vartheta(t)} = X$.

*We refer to the set of all ex-post behavioral efficient SCFs as the **ex-post behavioral efficient SCS** and denote it as EE.*

The EE SCS is non-empty whenever the ex-post choices are non-empty valued because a behaviorally efficient alternative exists at every type profile (de Clippel, 2014).

An SCF f is **ex-post Pareto efficient** (in the rational domain) if there is no $h \in H$ such that for some $t \in T$, we have $h(\vartheta(t)) P_{i,\vartheta(t)} f(\vartheta(t))$ for all $i \in N$.[15] We refer to the set of all ex-post Pareto efficient SCFs as the **ex-post Pareto efficient SCS** and denote it as $EXPO$.

The following lemma implies that ex-post behavioral efficiency extends ex-post Pareto efficiency to behavioral domains:

Lemma 1. *Given an ex-post environment, if individuals' ex-post choices are rational, then $EE = EXPO$.*

[15] This notion is the weak version of ex-post Pareto efficiency in Holmström and Myerson (1983).

Proof. Suppose individuals' ex-post choices are rational so that for all $i \in N$ and $t \in T$, there exists a complete, transitive, and reflexive preference relation $R_{i,\vartheta(t)}$ such that for any non-empty $S \subset X$, $x \in c_i^{\vartheta(t)}(S)$ if and only if $xR_{i,\vartheta(t)}y$ for all $y \in S$.

To see $EXPO \subset EE$, let f be ex-post Pareto efficient and define, for all i and θ, $Y_{i,\theta} = LCS_{i,\theta}(f(\theta)) = \{y \in X \mid f(\theta)R_{i,\theta}y\}$. Since $f(\theta) \in Y_{i,\theta}$ and every element of $Y_{i,\theta}$ is weakly worse than $f(\theta)$, it follows that $f(\vartheta(t)) \in c_i^{\vartheta(t)}(Y_{i,\vartheta(t)})$ for all i and t, establishing condition (i) of ex-post behavioral efficiency. Suppose, for contradiction, that there exist $\bar{t} \in T$ and $y \in X$ such that $y \notin Y_{i,\vartheta(\bar{t})}$ for all i. Then $yP_{i,\vartheta(\bar{t})}f(\vartheta(\bar{t}))$ for all i. Define $h : \Theta \to X$ by $h(\vartheta(\bar{t})) = y$ and $h(\theta') = f(\theta')$ for all $\theta' \neq \vartheta(\bar{t})$. This contradicts the ex-post Pareto efficiency of f, and hence condition (ii) holds.

To see why $EE \subset EXPO$, let f be ex-post behaviorally efficient but not ex-post Pareto efficient. Then there exist $h : \Theta \to X$ and $\bar{t} \in T$ such that $h(\vartheta(\bar{t}))P_{i,\vartheta(\bar{t})}f(\vartheta(\bar{t}))$ for all i. By condition (ii), $h(\vartheta(\bar{t})) \in Y_{j,\vartheta(\bar{t})}$ for some j, implying $f(\vartheta(\bar{t})) \notin c_j^{\vartheta(\bar{t})}(Y_{j,\vartheta(\bar{t})})$, a contradiction to (i) of ex-post behaviorally efficiency. ■

As we establish in our necessity result for ex-post behavioral implementation (Theorem 1), the existence of an ex-post consistent profile implies the quasi-ex-post incentive compatibility of the corresponding SCS (see Proposition 1). Therefore, quasi-ex-post incentive compatibility arises as a necessary condition for ex-post behavioral implementation. This is why we define a notion of ex-post behavioral incentive efficiency by restricting feasibility based on quasi-ex-post incentive compatibility (as in Holmström and Myerson (1983)):

Definition 6. *Given an ex-post environment, an SCF $f : \Theta \to X$ is **ex-post behavioral incentive efficient** if there is a profile of sets of alternatives $(Y_{i,\theta_{-i}})_{i \in N, \theta_{-i} \in \Theta_{-i}}$ such that*

(i) for all $i \in N$ and all $t \in T$, $f(\vartheta(t)) \in c_i^{\vartheta(t)}(Y_{i,\vartheta_{-i}(t_{-i})})$, and

(ii) for all $t \in T$, $\cup_{i \in N} Y_{i,\vartheta_{-i}(t_{-i})} = X$.

*We refer to the set of all ex-post behavioral incentive efficient SCFs as the **ex-post behavioral incentive efficient SCS** and denote it as EIE.*

The EIE SCS embeds *quasi–ex-post incentive compatibility* into the set of ex-post efficient (EE) social choice sets by requiring that the associated *implicit opportunity sets* be independent of individuals' private information. Formally, fix an SCF $f \in EIE$, a type profile $t \in T$, and an individual $i \in N$ and let $S = \left(Y^f_{i,\vartheta_{-i}(t_{-i})}\right)$. Then condition (i) of *ex–post behavioral incentive efficiency* implies that f satisfies quasi–ex-post incentive compatibility. Hence, every SCF in the EIE SCS is quasi–ex-post incentive compatible.

In contrast to our approach, the incentive compatibility notion in Holmström and Myerson involves the *interim* stage. In behavioral environments, its counterpart—quasi-interim incentive compatibility (Barlo & Dalkıran, 2023a)—is not implied by quasi-ex-post incentive compatibility.[16] Establishing a link requires an interim framework together with a condition that connects interim and ex-post choices. This is precisely what we do in Section 5, where we introduce *Property STP**. This property, which holds under rationality, requires that an act be chosen from a set of acts whenever, for every state, the alternative realized by that act is selected ex-post from the collection of alternatives supported by the acts in the set. Consequently, *ex–post behavioral incentive efficiency* is a refinement of Holmström and Myerson's ex-post incentive efficiency in rational domains.

Evidently, one cannot dispense with quasi-ex-post incentive compatibility, as it is a necessary condition for implementation in the EBE. Crucially, we show that no further restrictions are required to establish the implementability of the EIE SCS in EBE.

Proposition 5. *Suppose the ex-post environment is such that $n \geq 3$, ex-post choice incompatibility holds, and the EIE SCS is non-empty. Then, the EIE SCS is ex-post behavioral implementable.*

Proof. Let $f \in EIE$, $\theta_{-i} \in \Theta_{-i}$, and define $\mathbb{S} := (S_i(f,\theta_{-i}))_{i\in N,\ f\in F,\ \theta_{-i}\in\Theta_{-i}}$ by $S_i(f,\theta_{-i}) := Y^f_{i,\theta_{-i}}$ for all $i \in N$ as the profile of implicit opportunity sets associated with f as in Definition 6. Then, (i) of ex-post behavioral incentive efficiency implies (i) of ex-post consistency. Suppose for some $\tilde{f} \notin EIE$ and deception profile α, we have $\tilde{f}\circ\vartheta = f\circ\vartheta\circ\alpha$. Consider the profile of sets $(Y^f_{i,\alpha_{-i}(\theta_{-i})})_{i\in N}$. Observe that, by (ii) of ex-post behavioral efficiency of SCF f, we have for all $t \in T$, $\cup_{i\in N} Y^f_{i,\vartheta_{-i}(\alpha_{-i}(t_{-i}))} = X$. Therefore, as $\tilde{f} \notin EIE$, (i) of ex-post behavioral efficiency cannot hold for $\tilde{f}$, which means there are i^* and t^* such that $\tilde{f}(\vartheta(t^*)) \notin c_{i^*}^{\vartheta(t^*)}(Y^f_{i^*,\vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*}))})$ while $\tilde{f}(\vartheta(t^*)) = f(\vartheta(\alpha(t^*)))$. Because $Y^f_{i,\alpha_{-i}(\theta_{-i})} = S_i(f,\alpha_{-i}(\theta_{-i}))$ for all $i \in N$, all $\alpha \in \Lambda$, and all $\theta \in \Theta$, it follows that $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(S_{i^*}(f,\vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*}))))$, implying (ii) of ex-post consistency. Hence, $\mathbb{S}$ is ex-post consistent with the EIE SCS. The result follows from Theorem 2. ■

To complement our positive implementability result, we conclude this section with the following existence result:

Proposition 6. *Given an ex-post environment, if there is a quasi-ex-post incentive compatible SCF f such that for any $t \in T$, there is an individual $i^* \in N$ with $f(\tilde{\theta}_{i^*}, \vartheta_{-i^*}(t_{-i^*})) \in c_{i^*}^{(\tilde{\theta}_{i^*},\vartheta_{-i^*}(t_{-i^*}))}(X)$ for all $\tilde{\theta}_{i^*} \in \Theta_{i^*}$, then the EIE SCS is non-empty.*

[16] In the rational domain, quasi-interim incentive compatibility coincides with the interim incentive compatibility of Holmström and Myerson.

Proof. Let SCF f be quasi-ex-post incentive compatible and let the corresponding profile of sets of alternatives be $(S_{i,\theta})_{i\in N,\theta\in\Theta}$ where $S_{i,\vartheta(t)}$ is the set of alternatives associated with individual i, SCF f, and type profile t as specified in the definition of quasi-ex-post incentive compatibility. Recall that $f(\Theta_i, \vartheta_{-i}(t_{-i})) \subset S_{i,\vartheta(t)}$ for all $i \in N$ and $t \in T$. Let the implicit opportunity set profile $(Y_{i,\theta_{-i}})_{i\in N,\theta_{-i}\in\Theta_{-i}}$ be defined as follows: For all $t \in T$, $Y_{i,\vartheta_{-i}(t_{-i})} = S_{i,\vartheta(t)}$ for all $i \neq i^*(t)$, and $Y_{i^*,\vartheta_{-i^*(t)}(t_{-i^*(t)})} = X$ where $i^*(t)$ is the individual as described in the statement of the proposition associated with type profile t. Then, (i) of ex-post behavioral incentive efficiency holds as for all $t \in T$, $f(\tilde{\theta}_{i^*(t)}, \vartheta_{-i^*(t)}(t_{-i^*(t)})) \in c_{i^*(t)}^{(\tilde{\theta}_{i^*(t)},\vartheta_{-i^*(t)}(t_{-i^*(t)}))}(Y_{i^*(t),\vartheta_{-i^*(t)}(t_{-i^*(t)})})$ for all $\tilde{\theta}_{i^*(t)} \in \Theta_{i^*(t)}$; and for all $i \neq i^*(t)$, $f(\vartheta(t)) \in c_i^{\vartheta(t)}(S_{i,\vartheta(t)})$. Moreover, (ii) of ex-post behavioral efficiency holds as for all $t \in T$, $\cup_{i\in N} Y_{i,\vartheta_{-i}(t_{-i})} = X$ since $Y_{i^*(t),\vartheta_{-i^*(t)}(t_{-i^*(t)})} = X$. So, $f \in EIE$. ■

5 Interim and Ex-post Choices, and The Sure Thing Principle

To relate interim and ex-post choices, we present the following preliminaries:

For any individual $i \in N$, an *interim act* is $\mathbf{a}_i : T_{-i} \to X$, a function mapping T_{-i} into X. We denote the set of all interim acts of individual i by $\mathbf{A}_i$. The *image set associated with a set of acts* $\tilde{\mathbf{A}}_i \subset \mathbf{A}_i$ at t_{-i} equals $\tilde{\mathbf{A}}_i(t_{-i}) := \{x \in X \mid \mathbf{a}_i(t_{-i}) = x \text{ for some } \mathbf{a}_i \in \tilde{\mathbf{A}}_i\}$. Given $i \in N$, $t_i \in T_i$, and a non-empty subset of acts $\mathbf{S} \subset \mathbf{A}_i$, the *interim choice of individual i of type t_i from the set of acts* $\mathbf{S}$ is given by $\mathbf{C}_i^{t_i}(\mathbf{S}) \subset \mathbf{S}$. We wish to highlight that individuals' *beliefs* (assessments of others' types at the interim stage) are embedded into their interim choices.[17]

To see that beliefs are embedded in individuals' interim choices, consider the following example in which interim choices do not depend on the (ex-post) choice-state: Let $N = \{1, 2\}$, $X = \{x, y\}$, $T_i = \{t_i', t_i''\}$, and $\Theta_i = \{\bar{\theta}_i\}$ for both $i = 1, 2$. Even though there is one choice-type of both individuals, each individual has two types determining their beliefs. Our setup allows Individual 1 of type t_1' choosing only $\langle xx\rangle$ from the set of acts $\{\langle xx\rangle, \langle xy\rangle, \langle yx\rangle, \langle yy\rangle\}$ and Individual 1 of type t_1'' choosing only $\langle yy\rangle$ from the same set of acts where $\langle ab\rangle$ is the act $\mathbf{a}_1 : T_2 \to X$ with $\mathbf{a}_1(t_2') = a$ and $\mathbf{a}_1(t_2'') = b$.

[17] To illustrate this under rationality and Savage-Bayesian probabilistic sophistication, consider the situation when individuals' ex-post choices are captured by state-contingent utilities $(u_i(x \mid \theta))_{i\in N,\theta\in\Theta,x\in X}$ where $u_i(x \mid \theta)$ is the utility individual i obtains from alternative x at payoff-state θ. Under the standard Savage-Bayesian formulation, the interim beliefs $(\pi_i(t_{-i} \mid t_i))_{i\in N,t_i\in T_i,t_{-i}\in T_{-i}}$ emerge where $\pi_i(t_{-i} \mid t_i) \in [0, 1]$ denotes the belief of individual i of type t_i about the other individuals' type profile being t_{-i}. Then, we obtain the corresponding interim environment as follows: For any non-empty set of acts $\mathbf{S}$, the choice of individual i of type t_i equals

$$\mathbf{C}_i^{t_i}(\mathbf{S}) = \left\{ \mathbf{a} \in \mathbf{S} \;\middle|\; \begin{array}{c} \sum_{t_{-i}\in T_{-i}} \pi_i(t_{-i} \mid t_i)\, u_i(\mathbf{a}(t_{-i}) \mid \vartheta_i(t_i), \vartheta_{-i}(t_{-i})) \geq \\ \sum_{t_{-i}\in\Theta_{-i}} \pi_i(t_{-i} \mid t_i)\, u_i(\mathbf{b}(t_{-i}) \mid \vartheta_i(t_i), \vartheta_{-i}(t_{-i})) \text{ for all } \mathbf{b} \in \mathbf{S} \end{array} \right\}.$$

We impose *no restrictions* on interim choices as in our ex-post analysis. In particular, individuals' interim choices are not necessarily non-empty valued. The resulting interim environment $\mathcal{E}$ is common knowledge among the individuals.

An SCF $h : \Theta \to X$ induces an associated act that individual i of type t_i faces: $\mathbf{h}_{i,t_i} \in \mathbf{A}_i$ defined by $\mathbf{h}_{i,t_i}(t_{-i}) := h(\vartheta_i(t_i), \vartheta_{-i}(t_{-i}))$ for all $t_{-i} \in T_{-i}$.

Given a mechanism $\mu = (M, g)$ and a strategy profile σ, the set of acts individual i can unilaterally generate constitute *individual i's opportunity set of acts under μ for σ_{-i}*, which is defined as follows:

$$\mathbf{O}_i^{\mu}(\sigma_{-i}) := \{\mathbf{a}_i \in \mathbf{A}_i \mid \exists m_i \in M_i \text{ s.t. } \mathbf{a}_i(t_{-i}) = g(m_i, \sigma_{-i}(t_{-i})) \text{ for all } t_{-i} \in T_{-i}\}.$$

Given individuals' choices on interim acts, a natural equilibrium concept is as follows[18]: A strategy profile $\sigma^* = (\sigma_i^*)_{i \in N}$ is a ***behavioral interim equilibrium*** (BIE) of mechanism $\mu = (M, g)$ if for all $i \in N$ and all $t_i \in T_i$, $\mathbf{h}_{i,t_i}^* \in \mathbf{C}_i^{t_i}(\mathbf{O}_i^{\mu}(\sigma_{-i}^*))$, where $\mathbf{h}_{i,t_i}^*$ is the interim act induced by SCF h^* for individual i of type θ_i so that $h^* \circ \vartheta = g \circ \sigma^*$. Intuitively, σ^* is a BIE of μ if any individual i of any type t_i chooses the interim act generated by the prescribed action, $\sigma_i^*(t_i)$, from her opportunity set of acts corresponding to others' strategy profile σ_{-i}^*.

5.1 Property STP*

A natural way to link EBE and BIE involves relating ex-post choices of individuals to their interim choices via the following property, which is in the spirit of Savage (1972)'s sure-thing principle and Property STP introduced by de Clippel (2023).

Definition 7. *Given ex-post environment $\mathcal{E}^{\text{ep}}$, the associated interim environment $\mathcal{E}$ satisfies* **Property STP*** *if the following holds for each individual $i \in N$ and each of her type $t_i \in T_i$: for all non-empty sets of acts $\tilde{\mathbf{A}}_i \subset \mathbf{A}_i$ and all $\mathbf{a}_i \in \tilde{\mathbf{A}}_i$, if $\mathbf{a}_i(t_{-i}) \in c_i^{\vartheta(t)}(\tilde{\mathbf{A}}_i(t_{-i}))$ for all $t_{-i} \in T_{-i}$, then $\mathbf{a}_i \in \mathbf{C}_i^{t_i}(\tilde{\mathbf{A}}_i)$.*

Given an ex-post environment, its associated interim counterpart satisfies Property STP* if the following holds: For any individual i of any type t_i and any subset of her acts, $\tilde{\mathbf{A}}_i$, an act $\mathbf{a}_i \in \tilde{\mathbf{A}}_i$ is in the interim choice of i of type t_i from $\tilde{\mathbf{A}}_i$ whenever $\mathbf{a}_i$ is such that for any one of others' type profile t_{-i}, alternative $\mathbf{a}_i(t_{-i})$ is in i's ex-post choice from the set of alternatives $\tilde{\mathbf{A}}_i(t_{-i})$ (the image set associated with $\tilde{\mathbf{A}}_i$ at t_{-i}) at choice-state $\vartheta(t)$.[19]

[18] See Saran (2011) and Barlo and Dalkıran (2023a).

[19] An immediate implication of Property STP* is the following: If $\mathbf{a}_i(t_{-i}) \in c_i^{\vartheta(t_i, t_{-i})}(\tilde{\mathbf{A}}_i(t_{-i}))$ for all $t_{-i} \in T_{-i}$, then $\mathbf{a}_i \in \mathbf{C}_i^{t_i'}(\tilde{\mathbf{A}}_i)$ whenever $\vartheta(t_i', t_{-i}) = \vartheta(t_i, t_{-i})$ for all $t_{-i} \in T_{-i}$. This is why the

Under Property STP*, we obtain arguments similar to those of Bergemann and Morris (2008) and justify the use of EBE in behavioral domains: Every EBE of mechanism μ is a BIE of μ.[20] That is, Property STP* is *sufficient* for every EBE of a mechanism μ to be one of its BIE. In any EBE, the ex-post no-regret property holds, i.e., "no agent would like to change his message even if she were to know the true type profile of the remaining agents" (Bergemann & Morris, 2008). Therefore, EBE constitutes a BIE featuring the ex-post no-regret property under Property STP*. Furthermore, following Bergemann and Morris (2011), we show that any EBE at a permissible type space induces an outcome equivalent EBE and BIE at every permissible type space. See Appendix D for the formalities.

We note that Property STP* holds in the standard rational framework under Savage-Bayesian probabilistic sophistication. On the other hand, the minimax-regret preferences of Savage (1951) provide a setting in which the interim choices fail the IIA (and hence WARP), while Property STP* is satisfied. Thus, the minimax-regret setting delivers an interesting behavioral environment where the use of EBE is plausible.

In environments with the minimax-regret preferences, each type of each individual chooses the act that minimizes her maximum regret. The regret of individual i of type t_i from act $\mathbf{a}_i$ at type profile (t_i, t_{-i}) equals the difference between the ex-post payoff i obtains and her maximum ex-post payoff at this type profile, i.e., $\max_{\mathbf{a}'_i \in \mathbf{S}_i}(u_i(\mathbf{a}'_i(t_{-i}) \mid \vartheta(t))) - u_i(\mathbf{a}_i(t_{-i}) \mid \vartheta(t))$ where $u_i(x \mid \theta)$ denotes i's ex-post payoff from alternative x at choice-state θ. Hence, individual i of type t_i weakly prefers act $\mathbf{a}_i$ to act $\tilde{\mathbf{a}}_i$ in a given set of acts $\mathbf{S}_i$ if

$$\begin{aligned} &\max_{t_{-i} \in T_{-i}} \left[\max_{\mathbf{a}'_i \in \mathbf{S}_i} \left(u_i(\mathbf{a}'_i(t_{-i}) \mid \vartheta(t)) - u_i(\mathbf{a}_i(t_{-i}) \mid \vartheta(t)) \right) \right] \\ &\quad \leq \max_{t_{-i} \in T_{-i}} \left[\max_{\mathbf{a}''_i \in \mathbf{S}_i} \left(u_i(\mathbf{a}''_i(t_{-i}) \mid \vartheta(t)) - u_i(\tilde{\mathbf{a}}_i(t_{-i}) \mid \vartheta(t)) \right) \right]. \end{aligned} \tag{1}$$

Proposition 7. *If ex-post environment $\mathcal{E}^{\text{ep}}$ and the associated interim environment $\mathcal{E}$ are related via minimax-regret preferences, then Property STP* holds.*

Proof. Suppose ex-post environment $\mathcal{E}^{\text{ep}}$ and the associated interim environment $\mathcal{E}$ are related via minimax-regret preferences. That is, ex-post choices are represented by state-contingent utility functions —they are rational and hence satisfy the IIA. Moreover, the

interim choices specified in the example at the beginning of this section cannot arise under Property STP* whenever ex-post choices are non-empty valued: As $\bar{\theta}$ is the unique choice-state, $c_1^{\bar{\theta}}(\{x, y\})$ must either be $\{x\}$ or $\{y\}$ or $\{x, y\}$. In all these cases, the interim choices of Individual 1 of type t'_1 and t''_1 being only $\langle xx \rangle$ and only $\langle yy \rangle$, respectively, violates Property STP*.

[20] It is easy to see that in general, a BIE of a mechanism need not be one of its EBE regardless of whether or not Property STP* holds.

interim choices can be represented as follows: For any pair of acts $\mathbf{a}_i$ and $\tilde{\mathbf{a}}_i$ in a given set of acts $\mathbf{S}_i$, individual i of type t_i weakly prefers $\mathbf{a}_i$ to $\tilde{\mathbf{a}}_i$ in $\mathbf{S}_i$ if inequality (1) holds.

If for any individual i of type t_i, we have $\mathbf{a}_i^*(t'_{-i}) \in c_i^{\vartheta(t_i,t'_{-i})}(\tilde{\mathbf{A}}_i(t'_{-i}))$ for all $t'_{-i} \in T_{-i}$ for some non-empty set of interim acts $\tilde{\mathbf{A}}_i$, then $\mathbf{a}_i^* \in \mathbf{C}_i^{t_i}(\tilde{\mathbf{A}}_i)$; i.e., Property STP* holds. This follows from $\mathbf{a}_i^*$ minimizing the maximum regret: $u_i(\mathbf{a}_i^*(t'_{-i}) \mid (\vartheta_i(t_i), \vartheta_{-i}(t'_{-i}))) = \max_{x \in \tilde{\mathbf{A}}_i(t'_{-i})} u_i(x \mid (\vartheta_i(t_i), \vartheta_{-i}(t'_{-i})))$ for all t'_{-i}. So, for all t'_{-i}, $\max_{\mathbf{a}'_i \in \tilde{\mathbf{A}}_i}(u_i(\mathbf{a}'_i(t'_{-i}) \mid (\vartheta_i(t_i), \vartheta_{-i}(t'_{-i}))) - u_i(\mathbf{a}_i^*(t_{-i}) \mid (\vartheta_i(t_i), \vartheta_{-i}(t'_{-i})))) = 0$, i.e., i of type t_i's maximum regret from $\mathbf{a}_i^*$ at $(\vartheta_i(t_i), \vartheta_{-i}(t'_{-i}))$ is 0 for all $t'_{-i} \in T_{-i}$. ■

Therefore, in the case of minimax-regret preferences, Property STP* holds even though the interim choices are not necessarily rational (Saran, 2011; Barlo & Dalkıran, 2023a). We refer the reader to Example 1 [continued] (p. 23) for a formal demonstration of interim choices associated with minimax-regret preferences failing WARP.

Whether or not Property STP* is a *necessary* condition for every EBE of a mechanism being one of its BIE emerges as a natural question. Below, we show that the answer is negative by providing an example (inspired by de Clippel (2023)) that involves a mechanism where EBE and BIE coincide, but Property STP* fails to hold.

Example 3. Let $N = \{1,2\}$, $X = \{x,y,z\}$, $T_1 = \{t_1^1\}$, $T_2 = \{t_2^1, t_2^2\}$, $\Theta_1 = \{\theta_1^1\}$, $\Theta_2 = \{\theta_2^1, \theta_2^2\}$ where $\vartheta(t_i^j) = \theta_i^j$ for all $i,j \in N$. We denote the act, $\mathbf{a}_1 : T_2 \to X$, individual 1 faces by $\langle ab \rangle$ where $\mathbf{a}_1(t_2^1) = a$ and $\mathbf{a}_1(t_2^2) = b$ with $a,b \in \{x,y,z\}$. As there is only one type of individual 1, any act of individual 2 is an alternative, i.e., $\mathbf{A}_2 = X$.

Let the state-contingent payoffs and mechanism μ be as in Table 7. Suppose that both

Θ	(θ_1^1, θ_2^1)	(θ_1^1, θ_2^2)
$(u_1(x \mid \theta), u_2(x \mid \theta))$	$(2,2)$	$(0,1)$
$(u_1(y \mid \theta), u_2(y \mid \theta))$	$(1,1)$	$(1,2)$
$(u_1(z \mid \theta), u_2(z \mid \theta))$	$(1,0)$	$(1,0)$

		Ind. 2	
		L	R
Ind. 1	U	x	y
	D	y	x

Table 7: State-contingent payoffs and mechanism μ.

types of individual 2 are rational. The ex-post choices of individual 1 are rational, but at the interim stage, individual 1 has an aversion against uncertainty in the sense that she chooses an action that minimizes the maximum difference between her payoffs among all possible states with respect to the stage-contingent payoffs in Table 7: For any $\tilde{\mathbf{A}}_1 \subset \mathbf{A}_1$,

$$\mathbf{C}_1^{t_1}(\tilde{\mathbf{A}}_1) := \arg \min_{\mathbf{a}_1 \in \tilde{\mathbf{A}}_1} \left[\max_{t_2, t'_2 \in T_2} \left(u_1(\mathbf{a}_1(t_2) \mid (\theta_1^1, \vartheta_2(t_2))) - u_1(\mathbf{a}_1(t'_2) \mid (\theta_1^1, \vartheta_2(t'_2))) \right) \right].$$

To identify EBE of the mechanism in Table 7, consider the ex-post choices of individuals 1 and 2 given the stage-contingent payoffs in Table 7. Because $c_1^{\vartheta(t_1,t_2^1)}(\{x,y\}) = \{x\}$,

$c_1^{\vartheta(t_1,t_2^2)}(\{x,y\}) = \{y\}$, $c_2^{\vartheta(t_1,t_2^1)}(\{x,y\}) = \{x\}$, and $c_2^{\vartheta(t_1,t_2^2)}(\{x,y\}) = \{y\}$, there are two EBEs of mechanism μ both inducing SCF $\langle xy\rangle$: $\sigma^{(*)}$ and $\sigma^{(**)}$ where $\sigma_1^{(*)}(t_1) = U$, $\sigma_2^{(*)}(t_2^1) = L$, and $\sigma_2^{(*)}(t_2^2) = R$; $\sigma_1^{(**)}(t_1) = D$, $\sigma_2^{(**)}(t_2^1) = R$, and $\sigma_2^{(**)}(t_2^2) = L$.

Next, considering the interim choices of individual 1, we see that $\mathbf{C}_1^{t_1}(\{\langle xy\rangle, \langle yx\rangle\}) = \{\langle xy\rangle, \langle yx\rangle\}$ as the difference of individual 1's payoffs under $\langle xy\rangle$ and $\langle yx\rangle$ are both 1. Therefore, individual 1 is indifferent between choosing U or D in mechanism μ. On the other hand, because individual 2's interim choices over acts are such that $\mathbf{C}_2^{t_2^1}(\{\langle x\rangle, \langle y\rangle\}) = \{\langle x\rangle\}$ and $\mathbf{C}_2^{t_2^2}(\{\langle x\rangle, \langle y\rangle\}) = \{\langle y\rangle\}$, $\sigma^{(*)}$ and $\sigma^{(**)}$ are the only BIEs of mechanism μ.

Therefore, every EBE of μ is a BIE of μ and every BIE of μ is an EBE of μ.

Finally, we note that Property STP* fails. To see why, consider interim choices of individual 1 over $\tilde{\mathbf{A}}_1 = \{\langle xy\rangle, \langle yx\rangle, \langle zz\rangle\}$. We have $\mathbf{C}_1^{t_1}(\tilde{\mathbf{A}}_1) = \{\langle zz\rangle\}$ as the difference of individual 1's payoffs under $\langle zz\rangle$ is 0 whereas the differences of individual 1's payoffs under $\langle xy\rangle$ and $\langle yx\rangle$ are both 1. Further, $\tilde{\mathbf{A}}_1(t_2^1) = \tilde{\mathbf{A}}_1(t_2^2) = \{x,y,z\}$ and ex-post choices of individual 1 is such that $c_1^{\vartheta(t_1,t_2^1)}(\{x,y,z\}) = \{x\}$, $c_1^{\vartheta(t_1,t_2^2)}(\{x,y,z\}) = \{y\}$, but $\langle xy\rangle \notin \mathbf{C}_1^{t_1}(\tilde{\mathbf{A}}_1) = \{\langle zz\rangle\}$. Hence, Property STP* fails to hold.

Notwithstanding, we establish that ex-post behavioral implementability does not prevent the emergence of bad BIE regardless of whether or not Property STP* holds. To do that we use the minimax-regret setting and revisit Example 1.

Example 1 (continued). Recall that in this example, $N = \{1,2\}$, $X = \{x,y,z\}$, $T_i = \{t_i, t_i'\}$, $\Theta_i = \{\theta_i, \theta_i'\}$, $\vartheta_i(t_i) = \theta_i$, and $\vartheta_i(t_i') = \theta_i'$ for all $i = 1,2$. Consequently, there are four possible choice-states of the world, i.e., $\Theta = \{\theta_1\theta_2, \theta_1'\theta_2, \theta_1\theta_2', \theta_1'\theta_2'\}$. Table 8 details the rational interdependent ex-post preferences, and Table 9 specifies the corresponding payoffs for the minimax regret setting in which Property STP* holds. In this example,

$R_{1,(\theta_1,\theta_2)}$	$R_{2,(\theta_1,\theta_2)}$	$R_{1,(\theta_1',\theta_2)}$	$R_{2,(\theta_1',\theta_2)}$	$R_{1,(\theta_1,\theta_2')}$	$R_{2,(\theta_1,\theta_2')}$	$R_{1,(\theta_1',\theta_2')}$	$R_{2,(\theta_1',\theta_2')}$
x	x	z	z	z	z	y	y
z	z	x	x	y	y	z	z
y	y	y	y	x	x	x	x

Table 8: Ex-post preferences

	$u_{1,(\theta_1,\theta_2)}(\cdot)$	$u_{2,(\theta_1,\theta_2)}(\cdot)$	$u_{1,(\theta_1',\theta_2)}(\cdot)$	$u_{2,(\theta_1',\theta_2)}(\cdot)$	$u_{1,(\theta_1,\theta_2')}(\cdot)$	$u_{2,(\theta_1,\theta_2')}(\cdot)$	$u_{1,(\theta_1',\theta_2')}(\cdot)$	$u_{2,(\theta_1',\theta_2')}(\cdot)$
x	1	1	0	0	-1	-1	-1	-1
y	$-1+\eta$	-1	-1	-1	0	0	1	1
z	0	0	1	$1-\varepsilon$	1	$1-\tilde{\varepsilon}$	0	0

Table 9: Ex-post payoffs

the unique ex-post behavioral incentive efficient SCF f is given by $\langle xzzy\rangle$.

The interim choices fail WARP whenever $\eta \in (0,1)$: Consider individual 1 of type t_1 such that $\vartheta_1(t_1) = \theta_1$ and note that her choices from the choice sets $\{\langle xx\rangle, \langle yz\rangle, \langle zy\rangle\}$ (implying regret figures of $(0,2)$, $(2-\eta,0)$, and $(1,1)$, resp.) and $\{\langle yz\rangle, \langle zy\rangle\}$ (resulting in regret figures of $(1-\eta,0)$ and $(0,1)$, resp.) equal $\{\langle zy\rangle\}$ and $\{\langle yz\rangle\}$, respectively. In the above, an act that individual 1 of type t_1 faces is $\langle ab\rangle$ where a is the outcome of this act when the type of individual 2 is $\tilde{t}_2 \in T_2$ so that $\vartheta_2(\tilde{t}_2) = \theta_2$ while b corresponds to the outcome of the same act if the type of individual 2 is $\hat{t}_2 \in T_2$ so that $\vartheta_2(\hat{t}_2) = \theta_2'$.[21]

Recall that the direct mechanism given in Table 2 ex-post behavioral implements $F = \{f\}$. However, this mechanism has a bad BIE whenever $\min\{\varepsilon, \tilde{\varepsilon}\} = 0$ and $\eta < 1$.

To see this, let $\varepsilon = 0$, $\eta < 1$, and consider the strategy profile where both types of individual 1 claim to be of type t_1 while both types of individual 2 claim to be of type t_2', which we refer to as $\alpha^{(*)}$, i.e., $\alpha_1^{(*)}(t_1) = \alpha_1^{(*)}(t_1') = t_1$ and $\alpha_2^{(*)}(t_2) = \alpha_1^{(*)}(t_2') = t_2'$. Thus, SCF $\langle zzzz\rangle$ emerges as z is the resulting alternative at every choice-state under the direct mechanism. Therefore, each type of each individual obtains the act $\langle zz\rangle$ when they conform to this strategy profile. Let us consider the deviations of each type of each individual from this strategy profile one by one. $\underline{(1,t_1)}$: The maximum regret of $\langle zz\rangle$ is $\max\{1,0\} = 1$. When individual 1 of type t_1 deviates and claims to be of type t_1', she obtains $\langle yy\rangle$, which implies a maximum regret of $\max\{2-\eta,1\} = 2-\eta$. Therefore, individual 1 of type t_1 does not have any incentives to deviate. $\underline{(1,t_1')}$: The maximum regret of $\langle zz\rangle$ is $\max\{0,1\} = 1$. When individual 1 of type t_1' deviates to truthtelling, she obtains $\langle yy\rangle$, which implies a maximum regret of $\max\{2,0\} = 2$. Hence, individual 1 of type t_1' does not have any incentives to deviate as well. $\underline{(2,t_2)}$: The maximum regret of $\langle zz\rangle$ is $\max\{1,0\} = 1$. Individual 2 of type t_2 deviating to truthtelling delivers her the act $\langle xx\rangle$, which implies a maximum regret of $\max\{0,1-\varepsilon\} = 1-\varepsilon = 1$. Thus, individual 2 of type t_2 does not have any incentive to deviate as $\varepsilon = 0$. $\underline{(2,t_2')}$: The maximum regret of $\langle zz\rangle$ is $\max\{0,1\} = 1$. If individual 2 of type t_2' deviates and claims to be of type t_2, she obtains $\langle xx\rangle$, implying a maximum regret of $\max\{2,2\} = 2$. So, individual 2 of type t_2' is also not willing to deviate from this strategy profile. Ergo, this strategy profile is a bad BIE of the direct mechanism that induces the SCF $\langle zzzz\rangle \notin F$ when $\varepsilon = 0$. Similar arguments imply that the following strategy profile is a bad BIE inducing $\langle zzzz\rangle$, an SCF not aligned with ex-post incentive efficiency if $\tilde{\varepsilon} = 0$ and $\eta < 1$: Both types of individual 1 claim to be of type t_1' whereas both types of individual 2 claim to be of type t_2.

Now, we establish that any (direct or indirect) mechanism μ implementing SCS $F =$

[21] Recall that our interpretation of this example involves a headquarters (the planner) of a firm consisting of two subdivisions (individuals), which are located in two separate countries. The type of a subdivision reveals information about its country's state (ex-post choice-type) but not that of the other as well as beliefs about the other's type. In the current version of the example, each subdivision evaluates the interim acts via the minimax regret preferences (following 'the regret minimization framework,' popularized by Amazon CEO Jeff Bezos) after observing their own type. See https://www.linkedin.com/advice/0/what-benefits-drawbacks-regret-minimization-framework.

$\{\langle xzzy\rangle\}$ in EBE has a bad BIE that induces SCF $\langle zzzz\rangle$ whenever $\varepsilon = \tilde{\varepsilon} = 0$ and $\eta < 1$.[22] To see why let σ^f be an EBE of μ with $g \circ \sigma^f = f \circ \vartheta$ and consider $\alpha^{(*)}$. Then, $f \circ \vartheta \circ \alpha^{(*)} = \langle zzzz\rangle$, and $\langle xx\rangle, \langle zz\rangle \in \mathbf{O}_1^\mu(\sigma_2^f \circ \alpha_2^{(*)}) \subset \{\langle xx\rangle, \langle yy\rangle, \langle zz\rangle\}$, whereas $\langle yy\rangle, \langle zz\rangle \in \mathbf{O}_2^\mu(\sigma_1^f \circ \alpha_1^{(*)}) \subset \{\langle xx\rangle, \langle yy\rangle, \langle zz\rangle\}$. It follows from the above analysis of the direct mechanism that $\langle zz\rangle \in \mathbf{C}_1^{\tilde{t}_1}(\mathbf{O}_1^\mu(\sigma_2^f \circ \alpha_2^{(*)}))$ for all $\tilde{t}_1 \in T_1$ if $\mathbf{O}_1^\mu(\sigma_2^f \circ \alpha_2^{(*)}) = \{\langle xx\rangle, \langle zz\rangle\}$, and $\langle zz\rangle \in \mathbf{C}_2^{\tilde{t}_2}(\mathbf{O}_2^\mu(\sigma_1^f \circ \alpha_1^{(*)}))$ for all $\tilde{t}_2 = T_2$ if $\mathbf{O}_2^\mu(\sigma_1^f \circ \alpha_1^{(*)}) = \{\langle yy\rangle, \langle zz\rangle\}$. Thus, it suffices to show that $\langle zz\rangle \in \mathbf{C}_i^{\tilde{t}_i}(\{\langle xx\rangle, \langle yy\rangle, \langle zz\rangle\})$ for all $\tilde{t}_i \in T_i$ and all $i = 1, 2$. The maximum regret figures of $\langle xx\rangle$, $\langle yy\rangle$, and $\langle zz\rangle$ associated with the set $\{\langle xx\rangle, \langle yy\rangle, \langle zz\rangle\}$ are as follows: For $\underline{(1, t_1)}$, the respective maximum regret figures are 2, $2 - \eta$, and 1; for $\underline{(1, t_1')}$, they are 2, 2, and 1. For $\underline{(2, t_2)}$, the respective maximum regret figures are $1 - \varepsilon$, 2, and 1; for $\underline{(2, t_2')}$, they are 2, $1 - \tilde{\varepsilon}$, and 1. So, $\langle zz\rangle \in \mathbf{C}_i^{\tilde{t}_i}(\mathbf{O}_i^\mu(\sigma_j^f \circ \alpha_j^{(*)}))$ for all $\tilde{t}_i \in T_i$ and all $i, j = 1, 2$ with $i \neq j$ when $\varepsilon = \tilde{\varepsilon} = 0$ and $\eta < 1$. Hence, we conclude that $\sigma^f \circ \alpha^{(*)}$ is a BIE of μ sustaining SCF $\langle zzzz\rangle \notin F$.

de Clippel (2023) presents a serious *warning* for the use of ex-post behavioral equilibrium in environments that involve ex-post choices failing rationality: The failure of the IIA for ex-post choices is at odds with the plausibility of the ex-post behavioral equilibrium.[23]

The condition de Clippel (2023) analyzes, namely Property STP, is closely related to our Property STP* but restricted to probabilistic sophistication. Property STP is systematically violated when ex-post choices do not satisfy the IIA (and hence WARP). In Appendix C, we discuss situations in which a contradiction along the lines of de Clippel (2023) may emerge in our behavioral setting: To justify the use of EBE, one needs to dismiss two states that are perceived to be equivalent or the interim choices being unique up to the resulting equivalence classes (see Appendix C for further details).

We note that when ex-post choices satisfy the IIA, a contradiction à la de Clippel cannot arise (even if interim choices fail WARP). Indeed, minimax-regret preferences provide such a setting: the ex-post choices satisfy the IIA, but the interim choices do not.

5.2 An Indirect Approach via Property STP*

In this section, we provide an indirect derivation of *ex-post consistency* by linking ex-post and interim environments using only Property STP*. This approach allows us to

[22] In Appendix B, we characterize situations in which direct mechanisms ex-post behavioral implement given SCFs. In behavioral domains, direct mechanisms may lose their applicability not only because of the existence of bad BIE but also due to the possible failure of the revelation principle (Saran, 2011).

[23] Many interesting behavioral settings involve ex-post choices failing the IIA, e.g., the rational shortlist method of Manzini and Mariotti (2007); the choice under status-quo bias analyzed in Samuelson and Zeckhauser (1988), Masatlioglu and Ok (2014), and Dean et al. (2017); the choice with attraction effect as in Huber et al. (1982), de Clippel and Eliaz (2012), and Ok et al. (2015); committee choices with Condorcet cycles as in Hurwicz (1986); among others.

obtain our necessity and sufficiency results for ex-post behavioral implementation directly from the corresponding interim results established in Barlo and Dalkıran (2023a).

We begin by deriving ex-post consistency—the necessary condition for ex-post behavioral implementation—by relating it to *interim consistency*, which Barlo and Dalkıran (2023a) identify as a necessary condition for behavioral interim implementation.[24]

To this end, we associate each ex-post environment with a specific interim environment. Let $\mathcal{E}^{\text{ep}}$ be an ex-post environment specified by $(T, \vartheta) = (T_i, \vartheta_i)_{i \in N}$, where each individual $i \in N$ makes ex-post choices from subsets of alternatives according to the correspondence $c_i^{\vartheta(t)} : \mathcal{X} \to 2^X$. We construct the associated interim environment $\mathcal{E}^*$ as follows. For each individual i of type $t_i \in T_i$, interim choices from any set of acts $\tilde{\mathbf{A}}_i \subset \mathbf{A}_i$ are defined by

$$\mathbf{a}_i \in \mathbf{C}_i^{t_i}(\tilde{\mathbf{A}}_i) \quad \text{if and only if} \quad \mathbf{a}_i(t_{-i}) \in c_i^{(\vartheta_i(t_i), \vartheta_{-i}(t_{-i}))}\big(\tilde{\mathbf{A}}_i(t_{-i})\big) \text{ for all } t_{-i} \in T_{-i}. \tag{2}$$

In words, individual i chooses an act from $\tilde{\mathbf{A}}_i$ if and only if for every possible type profile of the other agents, the realization of that act is chosen from the corresponding image of $\tilde{\mathbf{A}}_i$ in the ex-post environment.

The interim choice correspondence $\mathbf{C}_i^{t_i}$ is empty-valued on $\tilde{\mathbf{A}}_i$ whenever no act in $\tilde{\mathbf{A}}_i$ satisfies the requirement in (2).

We emphasize that the association between the interim and ex-post environments, $\mathcal{E}^*$ and $\mathcal{E}^{\text{ep}}$, relies exclusively on Property STP*. As a consequence, the interim choices induced in $\mathcal{E}^*$ by (2) satisfy Property STP* by construction.

We now establish the equivalence between a mechanism's EBE in $\mathcal{E}^{\text{ep}}$ and its BIE in the associated interim environment $\mathcal{E}^*$.

Lemma 2. *Let μ be a mechanism and interim choices in $\mathcal{E}^*$ be as in (2) given the ex-post choices in $\mathcal{E}^{\text{ep}}$. Then, σ^* is an EBE of μ in $\mathcal{E}^{\text{ep}}$ if and only if σ^* is a BIE of μ in $\mathcal{E}^*$.*

Proof. σ^* is an EBE of μ in $\mathcal{E}^{\text{ep}}$ if and only if for all $t \in T$, $(g \circ \sigma^*)(t) \in c_i^{\vartheta(t)}(O_i^\mu(\sigma^*_{-i}(t_{-i})))$ for all $i \in N$. By (2), this is equivalent to $(g \circ \sigma^*)_{i,t_i} \in \mathbf{C}_i^{t_i}(\mathbf{O}_i^\mu(\sigma^*_{-i}))$ for all $i \in N$ and all $t_i \in T_i$, (i.e., σ^* is a BIE of μ in $\mathcal{E}^*$) since $\mathbf{O}_i^\mu(\sigma^*_{-i})(t_{-i}) = O_i^\mu(\sigma^*_{-i}(t_{-i}))$ and $(g \circ \sigma^*)_{i,t_i}$ is given by $(g \circ \sigma^*)_{i,t_i}(t_{-i}) = g \circ \sigma^*(t_i, t_{-i})$ for all $t_{-i} \in T_{-i}$. ■

Lemma 2 implies that an SCS F is implementable in EBE in $\mathcal{E}^{\text{ep}}$ if and only if it is implementable in BIE in $\mathcal{E}^*$ whenever interim choices in $\mathcal{E}^*$ are linked to the ex-post choices in $\mathcal{E}^{\text{ep}}$ via (2). Consequently, Proposition 8 below establishes that ex-post consistency

[24] We thank an anonymous reviewer for suggesting this approach. The reviewer correctly observed that transforming an ex-post environment into a suitably defined interim environment via a sure-thing–type condition yields equivalent results through an indirect route.

in an ex-post environment $\mathcal{E}^{\text{ep}}$ is equivalent to the interim consistency in the interim environment $\mathcal{E}^*$ associated via (2). As SCFs depend only on choice states in our current setup, the interim consistency of Barlo and Dalkıran (2023a) takes the following form:

Definition 8. *A profile of sets of acts* $(\mathbf{S}_i(f,\alpha_{-i}))_{i\in N,\ f\in F,\ \alpha_{-i}\in\Lambda_{-i}}$ *is* **interim consistent with SCS** F *if*

1. *it is closed under deception, i.e.,* $\mathbf{a}_i \in \mathbf{S}_i(f,\alpha_{-i})$ *implies* $\mathbf{a}_i\circ\tilde{\alpha}_{-i} \in \mathbf{S}_i(f,\alpha_{-i}\circ\tilde{\alpha}_{-i})$ *for all* $i\in N$, *all* $f\in F$, *and all* $\alpha_{-i},\tilde{\alpha}_{-i}\in\Lambda_{-i}$; *and*

2. *for every SCF* $f\in F$,

   (*i*) *for all* $i\in N$ *and all* $t_i\in T_i$ *such that* $\vartheta_i(t_i)=\theta_i$, $\mathbf{f}_{i,\theta_i}\in\mathbf{C}_i^{t_i}(\mathbf{S}_i(f,\alpha_{-i}^{\text{id}}))$, *and*

   (*ii*) *for any* $\alpha\in\Lambda$ *and* $\tilde{f}\notin F$ *with* $\tilde{f}\circ\vartheta = f\circ\vartheta\circ\alpha$, *there exist* $i^*\in N$ *and* $\theta_{i^*}^*\in\Theta_{i^*}$ *such that* $\mathbf{f}^{\alpha}_{i^*,\theta_{i^*}^*}\notin\mathbf{C}_{i^*}^{t_{i^*}}(\mathbf{S}_{i^*}(f,\alpha_{-i^*}))$ *for some* $t_{i^*}\in T_{i^*}$ *with* $\vartheta_{i^*}(t_{i^*})=\theta_{i^*}^*$,

   where acts $\mathbf{f}_{i,\theta_i}:T_{-i}\to X$ *and* $\mathbf{f}^{\alpha}_{i^*,\theta_{i^*}^*}:T_{-i^*}\to X$ *are defined as follows:* $\mathbf{f}_{i,\theta_i}(t_{-i}) = f(\theta_i,\vartheta_{-i}(t_{-i}))$ *for all* $t_{-i}\in T_{-i}$, *and* $\mathbf{f}^{\alpha}_{i^*,\theta_{i^*}^*}(t_{-i^*}) = f(\theta_{i^*}^*,\vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*})))$ *for all* $t_{-i^*}\in T_{-i^*}$ *and all* $\alpha_{-i^*}\in\Lambda_{-i^*}$.

Proposition 8. *Let interim choices in* $\mathcal{E}^*$ *be as in (2) given the ex-post choices in* $\mathcal{E}^{\text{ep}}$. *There is a profile of sets of acts interim consistent with SCS* F *in* $\mathcal{E}^*$ *if and only if there is a profile of sets of alternatives that is ex-post consistent with SCS* F *in* $\mathcal{E}^{\text{ep}}$.

Proof. Observe that the interim choices of individual i of type t_i defined by (2) in $\mathcal{E}^*$ depend only on the (ex-post) image sets of acts. That is, for any $i\in N$, $t_i\in T_i$, and any two sets of acts, $\tilde{\mathbf{A}}_i,\hat{\mathbf{A}}_i\subset\mathbf{A}_i$ with $\tilde{\mathbf{A}}_i(t_{-i}) = \hat{\mathbf{A}}_i(t_{-i})$ for all $t_{-i}\in T_{-i}$, we have $\mathbf{a}_i\in\mathbf{C}_i^{t_i}(\tilde{\mathbf{A}}_i)$ if and only if $\mathbf{a}_i\in\mathbf{C}_i^{t_i}(\hat{\mathbf{A}}_i)$. This implies that without a loss of generality, we may restrict attention to a setting where $\vartheta_i:T_i\to\Theta_i$ is a bijection for all $i\in N$.

($\Rightarrow$) Let $(\mathbf{S}_i(f,\alpha_{-i}))_{i\in N,\ f\in F,\ \alpha_{-i}\in\Lambda_{-i}}$ be a profile of sets of acts interim consistent with SCS F in $\mathcal{E}^*$. Define the profile of sets of alternatives $\mathbb{S} := (S_i(f,\theta_{-i}))_{i\in N,\ f\in F,\ \theta_{-i}\in\Theta_{-i}}$ as follows: $S_i(f,\theta_{-i}) := \mathbf{S}_i(f,\alpha_{-i}^{\text{id}})(t_{-i})$ for $t_{-i}\in T_{-i}$ such that $\vartheta_{-i}(t_{-i})=\theta_{-i}$.

Then, by 2.(*i*) of interim consistency, for all $i\in N$ and all $t_i\in T_i$ with $\vartheta_i(t_i)=\theta_i$, $\mathbf{f}_{i,\theta_i}\in\mathbf{C}_i^{t_i}(\mathbf{S}_i(f,\alpha_{-i}^{\text{id}}))$. By (2), this is equivalent to for all $i\in N$ and all $\theta_i\in\Theta_i$, $\mathbf{f}_{i,\theta_i}(t_{-i})\in c_i^{(\vartheta_i(t_i),\vartheta_{-i}(t_{-i}))}(\mathbf{S}_i(f,\alpha_{-i}^{\text{id}})(t_{-i}))$ for all $t_{-i}\in T_{-i}$. As for all $i\in N$ and all $t_i\in T_i$, $\vartheta_i(t_i)=\theta_i$, we have $f(\theta_i,\vartheta_{-i}(t_{-i}))\in c_i^{(\theta_i,\vartheta_{-i}(t_{-i}))}(\mathbf{S}_i(f,\alpha_{-i}^{\text{id}})(t_{-i}))$ for all $t_{-i}\in T_{-i}$. So, for all $i\in N$ and $\theta_i\in\Theta_i$, $f(\theta_i,\vartheta_{-i}(t_{-i})) = f(\theta)\in c_i^{\theta}(\mathbf{S}_i(f,\alpha_{-i}^{\text{id}})(t_{-i}))$. Hence, for all $i\in N$ and all $\theta\in\Theta$, $f(\theta)\in c_i^{\theta}(S_i(f,\theta_{-i}))$. Thus, for all $i\in N$ and all $t\in T$, $f(\vartheta(t))\in c_i^{\vartheta(t)}(S_i(f,\vartheta_{-i}(t_{-i})))$, i.e., $\mathbb{S} = (S_i(f,\theta_{-i}))_{i\in N,\ f\in F,\ \theta_{-i}\in\Theta_{-i}}$ satisfies (*i*) of ex-post consistency.

By 2.(ii) of interim consistency, for any $\alpha \in \Lambda$ and $\tilde{f} \notin F$ with $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$, there exists $i^* \in N$ and $\theta^*_{i^*} \in \Theta_{i^*}$ such that $\mathbf{f}^{\alpha}_{i^*,\theta^*_{i^*}} \notin \mathbf{C}^{t_{i^*}}_{i^*}(\mathbf{S}_{i^*}(f, \alpha_{-i^*}))$ for some $t_{i^*} \in T_{i^*}$ with $\vartheta_{i^*}(t_{i^*}) = \theta^*_{i^*}$, where $\mathbf{f}^{\alpha}_{i^*,\theta^*_{i^*}}(t_{-i^*}) = f(\theta^*_{i^*}, \vartheta_{-i^*}(\alpha_{-i^*}(t_{-i^*})))$ for all $t_{-i^*} \in T_{-i^*}$. By (2), this holds if and only if there is $\tilde{t}_{-i^*} \in T_{-i^*}$ such that $f(\theta^*_{i^*}, \vartheta_{-i^*}(\alpha_{-i^*}(\tilde{t}_{-i^*})))$ $\notin c_{i^*}^{(\theta^*_{i^*}, \vartheta_{-i^*}(\tilde{t}_{-i^*}))}(\mathbf{S}_{i^*}(f, \alpha_{-i^*})(\tilde{t}_{-i^*}))$. $\mathbf{S}_{i^*}(f, \alpha_{-i^*}) = \mathbf{S}_{i^*}(f, \alpha^{\mathrm{id}}_{-i^*} \circ \alpha_{-i^*})$ by closedness under deception, so, $\mathbf{S}_{i^*}(f, \alpha_{-i^*})(\tilde{t}_{-i^*}) = \mathbf{S}_{i^*}(f, \alpha^{\mathrm{id}}_{-i^*} \circ \alpha_{-i^*}(\tilde{t}_{-i^*}) = \mathbf{S}_{i^*}(f, \alpha^{\mathrm{id}}_{-i^*})(\alpha_{-i^*}(\tilde{t}_{-i^*}))$. Thus, $f(\theta^*_{i^*}, \vartheta_{-i^*}(\alpha_{-i^*}(\tilde{t}_{-i^*})))$ is not in $c_{i^*}^{(\theta^*_{i^*}, \vartheta_{-i^*}(\tilde{t}_{-i^*}))}(\mathbf{S}_{i^*}(f, \alpha^{\mathrm{id}}_{-i^*})(\alpha_{-i^*}(\tilde{t}_{-i^*})))$. Let $t^* = (t^*_{i^*}, t^*_{-i^*}) \in T$ be such that $\vartheta_{i^*}(\alpha_{i^*}(t^*_{i^*})) = \theta^*_{i^*}$ and $t^*_{-i^*} = \tilde{t}_{-i^*}$. Hence, $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*}))))$. In sum, for any $\alpha \in \Lambda$ and $\tilde{f} \notin F$ with $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$, there are $t^* \in T$ and $i^* \in N$ such that $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*}))))$ i.e., $\mathbb{S} = (S_i(f, \theta_{-i}))_{i \in N,\ f \in F,\ \theta_{-i} \in \Theta_{-i}}$ satisfies (ii) of ex-post consistency.

($\Leftarrow$) Let $\mathbb{S} := (S_i(f, \theta_{-i}))_{i \in N,\ f \in F,\ \theta_{-i} \in \Theta_{-i}}$ be a profile of sets of alternatives ex-post consistent with SCS F in $\mathcal{E}^{\mathrm{ep}}$. Define the profile of sets of acts $(\mathbf{S}_i(f, \alpha_{-i}))_{i \in N,\ f \in F,\ \alpha_{-i} \in \Lambda_{-i}}$ as follows: $\mathbf{S}_i(f, \alpha^{\mathrm{id}}_{-i}) := \{\mathbf{a}_i \in \mathbf{A}_i \mid \mathbf{a}_i(t_{-i}) \in S_i(f, \theta_{-i}) \text{ for all } t_{-i} \in T_{-i} \text{ with } \vartheta_{-i}(t_{-i}) = \theta_{-i}\}$; and for all $\alpha_{-i} \in \Lambda_{-i}$, let $\mathbf{S}_i(f, \alpha_{-i}) := \{\mathbf{a}_i \circ \alpha_{-i} \mid \mathbf{a}_i \in \mathbf{S}_i(f, \alpha^{\mathrm{id}}_{-i})\}$. That is, the acts in $\mathbf{S}_i(f, \alpha_{-i})$ are the acts in $\mathbf{S}_i(f, \alpha^{\mathrm{id}}_{-i})$ garbled with the deception α_{-i}. Observe that, by construction, $\mathbf{S}_i(f, \alpha^{\mathrm{id}}_{-i})(t_{-i}) = S_i(f, \theta_{-i})$ for every $t_{-i} \in T_{-i}$ such that $\vartheta_{-i}(t_{-i}) = \theta_{-i}$. Furthermore, the profile of sets of acts $(\mathbf{S}_i(f, \alpha_{-i}))_{i \in N,\ f \in F,\ \alpha_{-i} \in \Lambda_{-i}}$ is closed under deception since $\mathbf{a}_i \in \mathbf{S}_i(f, \alpha_{-i})$ implies $\mathbf{a}_i = \tilde{\mathbf{a}}_i \circ \alpha_{-i}$ for some $\tilde{\mathbf{a}}_i \in \mathbf{S}_i(f, \alpha^{\mathrm{id}}_{-i})$, which implies for any $\tilde{\alpha}_{-i} \in \Lambda_{-i}$, $\mathbf{a}_i \circ \tilde{\alpha}_{-i} = \tilde{\mathbf{a}}_i \circ \alpha_{-i} \circ \tilde{\alpha}_{-i} \in \mathbf{S}_i(f, \alpha_{-i} \circ \tilde{\alpha}_{-i})$ since $\tilde{\mathbf{a}}_i \in \mathbf{S}_i(f, \alpha^{\mathrm{id}}_{-i})$.

Next, observe that, by (i) of ex-post consistency, for all $f \in F$, all $i \in N$, and all $t \in T$, $f(\vartheta(t)) \in c_i^{\vartheta(t)}(S_i(f, \vartheta_{-i}(t_{-i})))$. Hence, for all $i \in N$ and all $t_i \in T_i$ with $\vartheta_i(t_i) = \theta_i$, $\mathbf{f}_{i,\theta_i}(t_{-i}) \in c_i^{(\vartheta_i(t_i), \vartheta_{-i}(t_{-i}))}(\mathbf{S}_i(f, \alpha^{\mathrm{id}}_{-i})(t_{-i}))$ for all $t_{-i} \in T_{-i}$. By (2), this implies that for all $i \in N$ and all $t_i \in T_i$ with $\vartheta_i(t_i) = \theta_i$, $\mathbf{f}_{i,\theta_i} \in \mathbf{C}^{t_i}_i(\mathbf{S}_i(f, \alpha^{\mathrm{id}}_{-i}))$, which means $(\mathbf{S}_i(f, \alpha_{-i}))_{i \in N,\ f \in F,\ \alpha_{-i} \in \Lambda_{-i}}$ satisfies 2.(i) of interim consistency.

Finally, by (ii) of ex-post consistency, for any $\alpha \in \Lambda$ and $\tilde{f} \notin F$ with $\tilde{f} \circ \vartheta = f \circ \vartheta \circ \alpha$, there are $t^* \in T$ and $i^* \in N$ such that $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*}))))$. This implies, by (2), that there are $i^* \in N$ and $\theta^*_{i^*} \in \Theta_{i^*}$ such that $\mathbf{f}^{\alpha}_{i^*,\theta^*_{i^*}} \notin \mathbf{C}^{t_{i^*}}_{i^*}(\mathbf{S}_{i^*}(f, \alpha_{-i^*}))$ for some $t_{i^*} \in T_{i^*}$ with $\vartheta_{i^*}(t_{i^*}) = \theta^*_{i^*}$ since $\mathbf{S}_{i^*}(f, \alpha_{-i^*})(t^*_{-i^*}) = S_{i^*}(f, \vartheta_{-i^*}(\alpha_{-i^*}(t^*_{-i^*}))$ by construction. Thus, $(\mathbf{S}_i(f, \alpha_{-i}))_{i \in N,\ f \in F,\ \alpha_{-i} \in \Lambda_{-i}}$ satisfies 2.(ii) of interim consistency. ■

The above results show that our necessity result, Theorem 1, can be obtained as a corollary of the necessity result of Barlo and Dalkıran (2023a). This interim analysis, however, relies on substantially heavier machinery. Focusing directly on ex-post behavior allows us to isolate the essential restriction in a transparent manner, yielding a concise

and self-contained proof—presented in a single paragraph within the main text.

Ex-post consistency and interim consistency are fundamentally distinct concepts when interim choices are non-empty valued: In Example 1 [continued] (p. 23), using minimax regret preferences (which imply non-empty interim choices), we show that the unique ex-post behavioral incentive efficient SCS is implementable in EBE but not in BIE, due to the emergence of a "bad" BIE not aligned with ex-post behavioral incentive efficiency. In fact, as we show in this example, when $\varepsilon = \tilde{\varepsilon} = 0$ and $\eta < 1$, no direct or indirect mechanism implements the unique ex-post behavioral incentive efficient SCS in BIE. Indeed, there is no interim consistent profile of sets of acts in this example. However, the ex-post consistent profile of sets of alternatives exists as this SCS is implementable in EBE.

One can also obtain our sufficiency result, Theorem 2, from the sufficiency result of Barlo and Dalkıran (2023a). This is because if ex-post choice incompatibility condition holds in $\mathcal{E}^{\text{ep}}$, then interim choice incompatibility of Barlo and Dalkıran (2023a) holds in $\mathcal{E}^{*}$ for the interim choices defined by (2). We prove this result below after presenting the interim choice incompatibility condition of Barlo and Dalkıran (2023a) tailored to our current setting where SCFs depend only on choice states:

Definition 9. *The* **interim choice incompatibility** *holds in an interim environment $\mathcal{E}$ whenever the following holds: If for any SCF $h \in H$ and any $\bar{t} \in T$, a profile of sets of acts $(\tilde{\mathbf{A}}_i)_{i \in N}$ is such that*

(i) for all $i \in N$, $\mathbf{h}_{i,\vartheta_i(\bar{t}_i)} \in \tilde{\mathbf{A}}_i$ where $\mathbf{h}_{i,\vartheta_i(\bar{t}_i)}(t_{-i}) = h(\vartheta_i(\bar{t}_i), \vartheta_{-i}(t_{-i}))$ for all $t_{-i} \in T_{-i}$,

(ii) there is $\bar{j} \in N$ such that for all $i \in N \setminus \{\bar{j}\}$, $\tilde{\mathbf{A}}_i(\bar{t}_{-i}) = X$,

then there is $i^* \in N \setminus \{\bar{j}\}$ such that $\mathbf{h}_{i^*,\vartheta_i(\bar{t}_{i^*})} \notin \mathbf{C}_{i^*}^{\bar{t}_{i^*}}(\tilde{\mathbf{A}}_{i^*})$.

Proposition 9. *Given an ex-post environment $\mathcal{E}^{\text{ep}}$, let $\mathcal{E}^{*}$ be the interim environment associated via (2). Then, ex-post choice incompatibility in $\mathcal{E}^{\text{ep}}$ implies interim choice incompatibility in $\mathcal{E}^{*}$.*

Proof. Suppose ex-post choice incompatibility holds in $\mathcal{E}^{\text{ep}}$, i.e., for all $t \in T$, all $x \in X$, all $\bar{j} \in N$, there is $i^* \in N \setminus \{\bar{j}\}$ such that $x \notin c_{i^*}^{\vartheta(t)}(X)$. This implies the interim choices defined by (2) in $\mathcal{E}^{*}$ are such that the following hold: for all $t \in T$, all $x \in X$, and all $\bar{j} \in N$, there is $i^* \in N \setminus \{\bar{j}\}$ such that if $\hat{\mathbf{A}}_{i^*}(t_{-i^*}) = X$ for some $\hat{\mathbf{A}}_{i^*} \subset \mathbf{A}_{i^*}$, then i^* of type t_i^* cannot choose any $\mathbf{a}_{i^*} \in \hat{\mathbf{A}}_{i^*}$ with $\mathbf{a}_{i^*}(t_{-i^*}) = x$ from $\hat{\mathbf{A}}_{i^*}$. Let SCF $h \in H$, $\bar{t} \in T$, and a profile of sets of acts $(\tilde{\mathbf{A}}_i)_{i \in N}$ be such that (i) and (ii) in Definition 9 of interim choice incompatibility hold, and consider the same $\bar{t} \in T$, the same $\bar{j} \in N$. Suppose, for a contradiction, for all $i \in N \setminus \{\bar{j}\}$, we have $\mathbf{h}_{i,\vartheta_i(\bar{t}_i)} \in \mathbf{C}_i^{\bar{t}_i}(\tilde{\mathbf{A}}_i)$. Let $h(\vartheta(\bar{t})) = x$. Then,

for all $i \in N$, $\mathbf{h}_{i,\vartheta_i(\bar{t}_i)}(\bar{t}_{-i}) = x$. It follows from (i) of Definition 9 that for all $i \in N \setminus \{\bar{j}\}$, $\mathbf{h}_{i,\vartheta_i(\bar{t}_i)}(\bar{t}_{-i}) \in \tilde{\mathbf{A}}_i(\bar{t}_{-i})$. Because $\mathbf{h}_{i,\vartheta_i(\bar{t}_i)} \in \mathbf{C}_i^{\bar{t}_i}(\tilde{\mathbf{A}}_i)$, by (2), it follows that for all $i \in N \setminus \{\bar{j}\}$, $x \in c_i^{\vartheta(\bar{t})}(\tilde{\mathbf{A}}_i(\bar{t}_{-i}))$. Since, by (ii) of Definition 9, $\tilde{\mathbf{A}}_i(\bar{t}_{-i}) = X$ for all $i \in N \setminus \{\bar{j}\}$, it follows that for all $i \in N \setminus \{\bar{j}\}$, $x \in c_i^{\vartheta(\bar{t})}(X)$. But, this contradicts ex-post choice incompatibility since for $\bar{t}$, x and $\bar{j} \in N$, there does not exists i^* such that $x \notin c_{i^*}^{\vartheta(t)}(X)$. ■

6 Concluding Remarks

In this paper, we have studied ex-post behavioral implementation without requiring individuals' ex-post and interim choices to satisfy the weak axiom of revealed preferences. Our analysis offers novel insights into behavioral mechanism design, particularly in settings where information asymmetries are inescapable.

Our results can be viewed as the behavioral counterpart of Bergemann and Morris (2008), which investigates ex-post implementation in rational domains, and as the ex-post analogue of de Clippel (2014) and Barlo and Dalkıran (2023a), which study behavioral implementation under complete information and behavioral interim implementation under incomplete information without any ex-post considerations, respectively.

We establish necessary as well as sufficient conditions for the ex-post behavioral implementation of social choice rules. As an application, we adapt the ex-post incentive efficiency of Holmström and Myerson (1983) to behavioral domains and show that it is fully ex-post behavioral implementable under mild conditions.

We hope our findings contribute to a deeper understanding of behavioral implementation and its practical relevance in environments with asymmetric information.

Appendix

A An Expositional Example

The following example aims to illustrate how to work out whether or not a given SCS is ex-post behavioral implementable, where individuals' ex-post choices fail WARP.

Individuals' choices involve three types of behavioral biases: (1) *attraction effect*, (2) *status-quo bias*, and (3) *Condorcet cycles*. Indeed, the example demonstrates that one can achieve ex-post behavioral implementation with different behavioral biases in different states of the world. Two individuals, *Ann* and *Bob*, are to decide what type of energy to employ or jointly invest in, be it *coal* energy, *nuclear* energy, or *solar* energy. Thus, the grand set of alternatives is $X = \{coal, nuclear, solar\}$.[25] Suppose that $T_i = \{t_i', t_i''\}$ and $\Theta_i = \{\rho_i, \gamma_i\}$ with $\vartheta_i(t_i') = \rho_i$ and $\vartheta_i(t_i'') = \gamma_i$ for both $i \in \{A, B\}$. The ex-post choices of Ann and Bob at choice-state $\theta \in \Theta$ are described by $c_A^\theta : \mathcal{X} \to \mathcal{X}$, and $c_B^\theta : \mathcal{X} \to \mathcal{X}$. Table 10 pinpoints the specific choices where c stands for *coal*, n for *nuclear* power, and s for *solar* energy.

S	$c_A^{(\rho_A,\rho_B)}$	$c_B^{(\rho_A,\rho_B)}$	$c_A^{(\rho_A,\gamma_B)}$	$c_B^{(\rho_A,\gamma_B)}$	$c_A^{(\gamma_A,\rho_B)}$	$c_B^{(\gamma_A,\rho_B)}$	$c_A^{(\gamma_A,\gamma_B)}$	$c_B^{(\gamma_A,\gamma_B)}$
$\{c,n,s\}$	$\{n\}$	$\{s\}$	$\{n\}$	$\{n\}$	$\{n\}$	$\{c\}$	$\{c,s\}$	$\{n,s\}$
$\{c,n\}$	$\{n\}$	$\{n\}$	$\{n\}$	$\{n\}$	$\{n\}$	$\{c\}$	$\{n\}$	$\{c\}$
$\{c,s\}$	$\{c,s\}$	$\{s\}$	$\{s\}$	$\{s\}$	$\{c\}$	$\{c\}$	$\{c\}$	$\{s\}$
$\{n,s\}$	$\{n\}$	$\{s\}$	$\{s\}$	$\{s\}$	$\{n,s\}$	$\{n,s\}$	$\{s\}$	$\{s\}$

Table 10: Ex-post choices of Ann and Bob.

At choice-state (ρ_A, ρ_B), Ann's ex-post choices can be rationalized by the preference relation $n \succ_A c \sim_A s$, and Bob's by $s \succ_B n \succ_B c$. The identical ex-post choices of Ann and Bob at (ρ_A, γ_B) can be explained by the *attraction effect*, one of the commonly observed behavioral biases. On the other hand, at state (γ_A, ρ_B), Bob's ex-post choices can be rationalized by the preference relation $c \succ_B s \sim_B n$, whereas Ann's ex-post choices feature a *status-quo bias* where the status-quo is c. Finally, at choice-state (γ_A, γ_B), the ex-post choices of Ann and Bob can be explained by 'groups as participants:' As the aggregation of individual preferences, neither of these ex-post choices can be rationalized by a complete and transitive preference relation because they violate both the IIA and Sen's β, and Ann's ex-post choices lead to a Condorcet cycle.

[25] Ann and Bob can also be interpreted as regions A and B within the same legislation, such as two states in the U.S. or two countries in the E.U. In his Nobel Prize Lecture "Mechanism Design: How to Implement Social Goals" (December 8, 2007), Eric Maskin provides an example in which an energy authority "is charged with choosing the type of energy to be used by Alice and Bob."

Our model allows individuals' ex-post choices to be *interdependent*: between choice-states (ρ_A, ρ_B) and (ρ_A, γ_B), Ann's private information (type t') does not change; yet, the ex-post choice behavior of Ann is not identical at these two states.

The planner aims to implement the SCS, $F = \{f, f'\}$, described in Table 11. We note that SCF f is ex-post behavioral incentive efficient while SCF f' is quasi-ex-post incentive compatible but not ex-post behavioral incentive efficient.[26]

	(ρ_A, ρ_B)	(ρ_A, γ_B)	(γ_A, ρ_B)	(γ_A, γ_B)
f	n	n	n	s
f'	s	s	c	c

Table 11: The social choice set F for Ann and Bob.

We now show that the following indirect mechanism $\mu = (M, g)$ fully ex-post behavioral implements the SCS $F = \{f, f'\}$, described in Table 11: $M_A = \{U, M, D\}$ and $M_B = \{L, M, R\}$; $g : M \to X$ is described in Table 12.

		Bob		
		L	M	R
Ann	U	n	c	n
	M	c	s	c
	D	n	s	s

Table 12: The mechanism μ for Ann and Bob.

Claim 6. *The strategy profiles* $\sigma'^* = (\sigma'^*_A, \sigma'^*_B)$, $\sigma''^* = (\sigma''^*_A, \sigma''^*_B)$, *and* $\sigma'''^* = (\sigma'''^*_A, \sigma'''^*_B)$ *described below are the only EBE of the mechanism* $\mu = (M, g)$, *where the outcomes generated under* σ''^* *and* σ'''^* *are equivalent, i.e.,* $g(\sigma''^*(t)) = g(\sigma'''^*(t))$ *for each* $t \in T$.

$$\sigma'^* : \quad \sigma'^*_A(t'_A) = U \quad \sigma'^*_A(t''_A) = D \quad \text{and} \quad \sigma'^*_B(t'_B) = L \quad \sigma'^*_B(t''_B) = R,$$
$$\sigma''^* : \quad \sigma''^*_A(t'_A) = D \quad \sigma''^*_A(t''_A) = U \quad \text{and} \quad \sigma''^*_B(t'_B) = M \quad \sigma''^*_B(t''_B) = M,$$
$$\sigma'''^* : \quad \sigma'''^*_A(t'_A) = M \quad \sigma'''^*_A(t''_A) = U \quad \text{and} \quad \sigma'''^*_B(t'_B) = M \quad \sigma'''^*_B(t''_B) = M.$$

Table 13 summarizes the EBE outcomes of μ where message profiles corresponding to σ'^* are depicted with circles while those associated with σ''^* are indicated with squares and those corresponding to σ'''^* with diamonds in the corresponding cells.

We observe that $g(\sigma'^*(t)) = f(t)$ for each $t \in T$, and $g(\sigma''^*(t)) = g(\sigma'''^*(t)) = f'(t)$ for each $t \in T$. Thus, μ fully ex-post behavioral implements the SCS F.

[26] To see why SCF f is ex-post behavioral incentive efficient, observe that sets $Y_{A,\rho_B} = Y_{A,\gamma_B} = \{c, n, s\}$, $Y_{B,\rho_A} = \{c, n\}$, and $Y_{B,\gamma_A} = \{n, s\}$ satisfy (i) and (ii) of Definition 6. To see why SCF f' is not ex-post behavioral incentive efficient, note that to satisfy (i) of Definition 6, we must have $Y_{A,\rho_B} = Y_{B,\rho_A} = \{c, s\}$. But, then (ii) of Definition 6 cannot hold since at (t'_A, t'_B), $Y_{A,\vartheta_B(t'_B)} \cup Y_{B,\vartheta_A(t'_A)} = \{c, s\} \neq X$.

$\vartheta(t'_A, t'_B) = (\rho_A, \rho_B)$

	L	M	R
U	$\textcircled{n}$	c	n
M	c	$\diamondsuit{s}$	c
D	n	$\boxed{s}$	s

$\vartheta(t'_A, t''_B) = (\rho_A, \gamma_B)$

	L	M	R
U	n	c	$\textcircled{n}$
M	c	$\diamondsuit{s}$	c
D	n	$\boxed{s}$	s

$\vartheta(t''_A, t'_B) = (\gamma_A, \rho_B)$

	L	M	R
U	n	$\diamondsuit{c}$	n
M	c	s	c
D	$\textcircled{n}$	s	s

$\vartheta(t''_A, t''_B) = (\gamma_A, \gamma_B)$

	L	M	R
U	n	$\diamondsuit{c}$	n
M	c	s	c
D	n	s	$\textcircled{s}$

Table 13: Ex-post equilibria and ex-post equilibrium outcomes of the mechanism.

Proof of Claim 6. We identify all EBE of $\mu = (M, g)$ described in Table 12 by a case-by-case analysis of what Ann plays when her type is t'_A.

Let σ^* be an EBE of $\mu = (M, g)$. There are three possible cases for $\sigma^*_A(t'_A)$ which can be either U or M or D.

Case 1. If $\sigma^*_A(t'_A) = U$: Then, $O^\mu_B(\sigma^*_A(t'_A)) = \{c, n\}$. At (ρ_A, ρ_B) and (ρ_A, γ_B), Bob chooses n from the set $\{c, n\}$. Thus, $\sigma^*_B(t'_B)$ and $\sigma^*_B(t''_B)$ must be either L or R.

Subcase 1.1. If $\sigma^*_B(t'_B) = L$ and $\sigma^*_B(t''_B) = L$: Then, $g(\sigma^*(t'_A, t'_B)) = n = g(\sigma^*(t'_A, t''_B))$. We have $O^\mu_A(\sigma^*_B(t'_B)) = O^\mu_A(\sigma^*_B(t''_B)) = \{c, n\}$. At (γ_A, ρ_B) and at (γ_A, γ_B), Ann chooses n from $\{c, n\}$ which implies $\sigma^*_A(t''_A)$ must be U or D. If $\sigma^*_A(t''_A) = U$, then $O^\mu_B(\sigma^*_A(t''_A)) = \{c, n\}$, which contradicts $\sigma^*_B(t''_B) = L$ as Bob chooses c from $\{c, n\}$ at (γ_A, γ_B). On the other hand, if $\sigma^*_A(t''_A) = D$, then $O^\mu_B(\sigma^*_A(t''_A)) = \{n, s\}$ but this contradicts with $\sigma^*_B(t''_B) = L$ as Bob chooses s from $\{n, s\}$ at (γ_A, γ_B). Thus, we cannot have $\sigma^*_B(t'_B) = L$ and $\sigma^*_B(t''_B) = L$.

Subcase 1.2. If $\sigma^*_B(t'_B) = L$ and $\sigma^*_B(t''_B) = R$: Then, $g(\sigma^*(t'_A, t'_B)) = n = g(\sigma^*(t'_A, t''_B))$. We have $O^\mu_A(\sigma^*_B(t'_B)) = \{c, n\}$ and $O^\mu_A(\sigma^*_B(t''_B)) = \{c, n, s\}$. At (γ_A, ρ_B), Ann chooses n from $\{c, n\}$, which implies $\sigma^*_A(t''_A)$ must be either U or D. At (γ_A, γ_B), Ann chooses c and s from $\{c, n, s\}$, which implies $\sigma^*_A(t''_A)$ must be M or D. So, $\sigma^*_A(t''_A) = D$.

Indeed, the following observations imply that our first EBE is σ'^* such that $\sigma'^*_A(\rho_A) = U$, $\sigma'^*_A(\gamma_A) = D$, and $\sigma'^*_B(\rho_B) = L$, $\sigma'^*_B(\gamma_B) = R$

$$
\begin{aligned}
\text{At } (\rho_A, \rho_B) \quad &: \quad n \in c^{(\rho_A,\rho_B)}_A(\{c,n\}) \implies g(\sigma'^*(t'_A, t'_B)) \in c^{(\rho_A,\rho_B)}_A(O^\mu_A(\sigma'^*_B(t'_B))),\\
&\qquad n \in c^{(\rho_A,\rho_B)}_B(\{c,n\}) \implies g(\sigma'^*(t'_A, t'_B)) \in c^{(\rho_A,\rho_B)}_B(O^\mu_B(\sigma'^*_A(t'_A))).\\
\text{At } (\rho_A, \gamma_B) \quad &: \quad n \in c^{(\rho_A,\gamma_B)}_A(\{c,n,s\}) \implies g(\sigma'^*(t'_A, t''_B)) \in c^{(\rho_A,\gamma_B)}_A(O^\mu_A(\sigma'^*_B(t''_B))),\\
&\qquad n \in c^{(\rho_A,\gamma_B)}_B(\{c,n\}) \implies g(\sigma'^*(t'_A, t''_B)) \in c^{(\rho_A,\gamma_B)}_B(O^\mu_B(\sigma'^*_A(t'_A))).\\
\text{At } (\gamma_A, \rho_B) \quad &: \quad n \in c^{(\gamma_A,\rho_B)}_A(\{c,n\}) \implies g(\sigma'^*(t''_A, t'_B)) \in c^{(\gamma_A,\rho_B)}_A(O^\mu_A(\sigma'^*_B(t'_B))),\\
&\qquad n \in c^{(\gamma_A,\rho_B)}_B(\{n,s\}) \implies g(\sigma'^*(t''_A, t'_B)) \in c^{(\gamma_A,\rho_B)}_B(O^\mu_B(\sigma'^*_A(t''_A))).\\
\text{At } (\gamma_A, \gamma_B) \quad &: \quad s \in c^{(\gamma_A,\gamma_B)}_A(\{c,n,s\}) \implies g(\sigma'^*(t''_A, t''_B)) \in c^{(\gamma_A,\gamma_B)}_A(O^\mu_A(\sigma'^*_B(t''_B))),\\
&\qquad s \in c^{(\gamma_A,\gamma_B)}_B(\{n,s\}) \implies g(\sigma'^*(t''_A, t''_B)) \in c^{(\gamma_A,\gamma_B)}_B(O^\mu_B(\sigma'^*_A(t''_A))).
\end{aligned}
$$

Subcase 1.3. If $\sigma_B^*(t'_B) = R$ and $\sigma_B^*(t''_B) = L$: Then, $g(\sigma^*(t'_A, t'_B)) = n = g(\sigma^*(t'_A, t''_B))$. We have $O_A^\mu(\sigma_B^*(t'_B)) = \{c, n, s\}$ and $O_A^\mu(\sigma_B^*(t''_B)) = \{c, n\}$. At (γ_A, ρ_B), Ann chooses n from $\{c, n, s\}$, which implies $\sigma_A^*(t''_A)$ must be U. On the other hand, at (γ_A, γ_B), Ann chooses n from $\{c, n\}$, which implies $\sigma_A^*(t''_A)$ must be U or D. Therefore, we must have $\sigma_A^*(t''_A) = U$. This implies $O_B^\mu(\sigma_A^*(t''_A)) = \{c, n\}$. But, at (γ_A, ρ_B), Bob chooses c from $\{c, n\}$ even though it would be $g(\sigma^*(t''_A, t'_B)) = n$, a contradiction. Hence, we cannot have $\sigma_B^*(t'_B) = R$ and $\sigma_B^*(t''_B) = L$.

Subcase 1.4. If $\sigma_B^*(t'_B) = R$ and $\sigma_B^*(t''_B) = R$: Then, $g(\sigma^*(t'_A, t'_B)) = n = g(\sigma^*(t'_A, t''_B))$. We have $O_A^\mu(\sigma_B^*(t'_B)) = O_A^\mu(\sigma_B^*(t''_B)) = \{c, n, s\}$. At (γ_A, ρ_B), Ann chooses n from $\{c, n, s\}$, which implies $\sigma_A^*(t''_A)$ must be U. But, at (γ_A, γ_B), Ann chooses c and s from $\{c, n, s\}$, which implies $\sigma_A^*(t''_A)$ must be either M or D, a contradiction. Therefore, we cannot have $\sigma_B^*(t'_B) = R$ and $\sigma_B^*(t''_B) = R$.

Case 2. If $\sigma_A^*(t'_A) = M$: Then, $O_B^\mu(\sigma_A^*(t'_A)) = \{c, s\}$. At (ρ_A, ρ_B) and (ρ_A, γ_B), Bob chooses s from the set $\{c, s\}$. Therefore, $\sigma_B^*(t'_B)$ and $\sigma_B^*(t''_B)$ must both be M. Then, $O_A^\mu(\sigma_B^*(t'_B)) = O_A^\mu(\sigma_B^*(t''_B)) = \{c, s\}$. At (γ_A, ρ_B) and (γ_A, γ_B) Ann chooses c from the set $\{c, s\}$, which implies it must be that $\sigma_A^*(t''_A) = U$.

Then, the following observations imply that our second EBE is σ'''^* such that $\sigma_A'''^*(t'_A) = M$, $\sigma_A'''^*(t''_A) = U$, and $\sigma_B'''^*(t'_B) = M$, $\sigma_B'''^*(t''_B) = M$

$$
\begin{aligned}
\text{At } (\rho_A, \rho_B) \quad : \quad & s \in c_A^{(\rho_A,\rho_B)}(\{c, s\}) \implies g(\sigma'''^*(t'_A, t'_B)) \in c_A^{(\rho_A,\rho_B)}(O_A^\mu(\sigma_B'''^*(t'_B))), \\
& s \in c_B^{(\rho_A,\rho_B)}(\{c, s\}) \implies g(\sigma'''^*(t'_A, t'_B)) \in c_B^{(\rho_A,\rho_B)}(O_B^\mu(\sigma_A'''^*(t'_A))). \\
\text{At } (\rho_A, \gamma_B) \quad : \quad & s \in c_A^{(\rho_A,\gamma_B)}(\{c, s\}) \implies g(\sigma'''^*(t'_A, t''_B)) \in c_A^{(\rho_A,\gamma_B)}(O_A^\mu(\sigma_B'''^*(t''_B))), \\
& s \in c_B^{(\rho_A,\gamma_B)}(\{c, s\}) \implies g(\sigma'''^*(t'_A, t''_B)) \in c_B^{(\rho_A,\gamma_B)}(O_B^\mu(\sigma_A'''^*(t'_A))). \\
\text{At } (\gamma_A, \rho_B) \quad : \quad & c \in c_A^{(\gamma_A,\rho_B)}(\{c, s\}) \implies g(\sigma'''^*(t''_A, t'_B)) \in c_A^{(\gamma_A,\rho_B)}(O_A^\mu(\sigma_B'''^*(t'_B))), \\
& c \in c_B^{(\gamma_A,\rho_B)}(\{c, n\}) \implies g(\sigma'''^*(t''_A, t'_B)) \in c_B^{(\gamma_A,\rho_B)}(O_B^\mu(\sigma_A'''^*(t''_A))). \\
\text{At } (\gamma_A, \gamma_B) \quad : \quad & c \in c_A^{(\gamma_A,\gamma_B)}(\{c, s\}) \implies g(\sigma'''^*(t''_A, t''_B)) \in c_A^{(\gamma_A,\gamma_B)}(O_A^\mu(\sigma_B'''^*(t''_B))), \\
& c \in c_B^{(\gamma_A,\gamma_B)}(\{c, n\}) \implies g(\sigma'''^*(t''_A, t''_B)) \in c_B^{(\gamma_A,\gamma_B)}(O_B^\mu(\sigma_A'''^*(t''_A))).
\end{aligned}
$$

Case 3. If $\sigma_A^*(t'_A) = D$: Then, $O_B^\mu(\sigma_A^*(t'_A)) = \{n, s\}$. At (ρ_A, ρ_B) and (ρ_A, γ_B), Bob chooses s from the set $\{n, s\}$. Therefore, $\sigma_B^*(t'_B)$ and $\sigma_B^*(t''_B)$ must be either M or R.

Subcase 3.1. If $\sigma_B^*(t'_B) = M$ and $\sigma_B^*(t''_B) = M$: So, $g(\sigma^*(t'_A, t'_B)) = s = g(\sigma^*(t'_A, t''_B))$. We have $O_A^\mu(\sigma_B^*(t'_B)) = O_A^\mu(\sigma_B^*(t''_B)) = \{c, s\}$. At (γ_A, ρ_B) and (γ_A, γ_B), Ann chooses c from $\{c, s\}$, which implies it must be $\sigma_A^*(t''_A) = U$.

Indeed, the following observations imply that our third EBE is σ''^* such that $\sigma_A''^*(\rho_A) = D$, $\sigma_A''^*(\gamma_A) = U$, and $\sigma_B''^*(\rho_B) = M$, $\sigma_B''^*(\gamma_B) = M$.

$$
\begin{aligned}
\text{At } (\rho_A,\rho_B) \quad &: \quad s \in c_A^{(\rho_A,\rho_B)}(\{c,s\}) \implies g(\sigma''^*(t'_A,t'_B)) \in c_A^{(\rho_A,\rho_B)}(O_A^\mu(\sigma''^*_B(t'_B))),\\
&\qquad s \in c_B^{(\rho_A,\rho_B)}(\{n,s\}) \implies g(\sigma''^*(t'_A,t'_B)) \in c_B^{(\rho_A,\rho_B)}(O_B^\mu(\sigma''^*_A(t'_A))).\\
\text{At } (\rho_A,\gamma_B) \quad &: \quad s \in c_A^{(\rho_A,\gamma_B)}(\{c,s\}) \implies g(\sigma''^*(t'_A,t''_B)) \in c_A^{(\rho_A,\gamma_B)}(O_A^\mu(\sigma''^*_B(t''_B))),\\
&\qquad s \in c_B^{(\rho_A,\gamma_B)}(\{n,s\}) \implies g(\sigma''^*(t'_A,t''_B)) \in c_B^{(\rho_A,\gamma_B)}(O_B^\mu(\sigma''^*_A(t'_A))).\\
\text{At } (\gamma_A,\rho_B) \quad &: \quad c \in c_A^{(\gamma_A,\rho_B)}(\{c,s\}) \implies g(\sigma''^*(t''_A,t'_B)) \in c_A^{(\gamma_A,\rho_B)}(O_A^\mu(\sigma''^*_B(t'_B))),\\
&\qquad c \in c_B^{(\gamma_A,\rho_B)}(\{c,n\}) \implies g(\sigma''^*(t''_A,t'_B)) \in c_B^{(\gamma_A,\rho_B)}(O_B^\mu(\sigma''^*_A(t''_A))).\\
\text{At } (\gamma_A,\gamma_B) \quad &: \quad c \in c_A^{(\gamma_A,\gamma_B)}(\{c,s\}) \implies g(\sigma''^*(t''_A,t''_B)) \in c_A^{(\gamma_A,\gamma_B)}(O_A^\mu(\sigma''^*_B(t''_B))),\\
&\qquad c \in c_B^{(\gamma_A,\gamma_B)}(\{c,n\}) \implies g(\sigma''^*(t''_A,t''_B)) \in c_B^{(\gamma_A,\gamma_B)}(O_B^\mu(\sigma''^*_A(t''_A))).
\end{aligned}
$$

Subcase 3.2. If $\sigma_B^*(t'_B) = M$ and $\sigma_B^*(t''_B) = R$: So, $g(\sigma^*(t'_A,t'_B)) = s = g(\sigma^*(t'_A,t''_B))$. We have $O_A^\mu(\sigma_B^*(t'_B)) = \{c,s\}$ and $O_A^\mu(\sigma_B^*(t''_B)) = \{c,n,s\}$. At (γ_A,ρ_B), Ann chooses c from $\{c,s\}$, which implies $\sigma_A^*(t''_A)$ must be U. On the other hand, at (γ_A,γ_B), Ann chooses c and s from $\{c,n,s\}$, which implies $\sigma_A^*(t''_A)$ must be M or D, a contradiction. Hence, we cannot have $\sigma_B^*(t'_B) = M$ and $\sigma_B^*(t''_B) = R$.

Subcase 3.3. If $\sigma_B^*(t'_B) = R$ and $\sigma_B^*(t''_B) = M$: So, $g(\sigma^*(t'_A,t'_B)) = s = g(\sigma^*(t'_A,t''_B))$. We have $O_A^\mu(\sigma_B^*(t'_B)) = \{c,n,s\}$ and $O_A^\mu(\sigma_B^*(t''_B)) = \{c,s\}$. At (γ_A,ρ_B), Ann chooses n from $\{c,n,s\}$, and at (γ_A,γ_B), Ann chooses c from $\{c,s\}$. They both imply we must have $\sigma_A^*(t''_A) = U$. Thus, $O_B^\mu(\sigma_A^*(t''_A)) = \{c,n\}$. But, at (γ_A,ρ_B), Bob chooses c from $\{c,n\}$ even though it would be $g(\sigma^*(t''_A,t'_B)) = n$, a contradiction. So, we cannot have $\sigma_B^*(t'_B) = R$ and $\sigma_B^*(t''_B) = M$.

Subcase 3.4. If $\sigma_B^*(t'_B) = R$ and $\sigma_B^*(t''_B) = R$: So, $g(\sigma^*(t'_A,t'_B)) = s = g(\sigma^*(t'_A,t''_B))$. We have $O_A^\mu(\sigma_B^*(t'_B)) = O_A^\mu(\sigma_B^*(t''_B)) = \{c,n,s\}$. At (γ_A,ρ_B), Ann chooses n from $\{c,n,s\}$, which implies $\sigma_A^*(t''_A)$ must be U. On the other hand, at (γ_A,γ_B), Ann chooses c,s from $\{c,n,s\}$, which implies $\sigma_A^*(t''_A)$ must be M or D, a contradiction. Thus, we cannot have $\sigma_B^*(t'_B) = R$ and $\sigma_B^*(t''_B) = R$ as well. ■

Finally, we observe that the revelation principle for partial ex-post behavioral implementation fails in our example.[27] Consider the SCF f given in Table 11. The mechanism μ in Table 12 possesses an EBE sustaining f. Hence, the mechanism μ partially ex-post behavioral implements f. However, the corresponding direct mechanism, μ^f, given in Table 14, fails to partially ex-post behavioral implement f truthfully as truthful revelation is not an EBE of μ^f: At state (t'_A,t''_B), reporting truthfully delivers n, but Ann's opportunity

[27]The failure of the revelation principle on behavioral domains is first documented by Saran (2011) for the concept of BIE. That study establishes that *weak contraction consistency*, a condition implied by the IIA, is sufficient for the revelation principle. Our example shows that the revelation also fails for the concept of EBE, which is essentially due to the failure of IIA in ex-post choices.

		Bob	
		t'_B	t''_B
Ann	t'_A	n	n
	t''_A	n	s

Table 14: The direct mechanism μ^f.

set at the corresponding choice-state (ρ_A, γ_B) is $\{n, s\}$ and $n \notin C_A^{(\rho_A,\gamma_B)}(\{n, s\}) = \{s\}$.

B Direct Mechanisms

The appeal of direct mechanisms in the mechanism design literature leads us to the following analysis, in which we focus on SCFs instead of SCSs (since direct mechanisms cannot coordinate selections of SCFs from an SCS). In Theorem 3, we present the resulting characterization of ex-post behavioral implementation of an SCF via its direct mechanism. Moreover, we provide a *second* characterization, which is akin to Bergemann and Morris (2008, Proposition 1), using the following condition we borrow from that study: An SCF f is *full-range* if for all $x \in X$, all $i \in N$, and all $\theta_{-i} \in \Theta_{-i}$, there is $\theta_i \in \Theta_i$ with $f(\theta) = x$.

Theorem 3. *Given an ex-post environment, let $f : \Theta \to X$ be an SCF. Then,*

(*i*) *SCS $F = \{f\}$ is ex-post behavioral implementable by f's direct mechanism possessing a truthful EBE if and only if the profile of sets $(f(\Theta_i, \theta_{-i}))_{i\in N,\ \theta_{-i}\in\Theta_{-i}}$ is ex-post consistent with F.*

(*ii*) *If f is full-range, then SCS $F = \{f\}$ is ex-post implementable if and only if it is ex-post behavioral implementable via the direct mechanism of f.*

Proof. For the *necessity* of (*i*) of the theorem, suppose f is ex-post behavioral implementable by its direct mechanism $\mu^f = (T, g^f)$ with $g^f = f \circ \vartheta$. Let the truthful strategy profile α^{id} be an EBE, so $f \circ \vartheta = g^f \circ \alpha^{\mathrm{id}}$. Let $i \in N$ and $t \in T$. Then, $O_i^{\mu^f}(\alpha_{-i}^{\mathrm{id}}(t_{-i})) = f(\Theta_i, \vartheta_{-i}(t_{-i}))$ and $f(\vartheta_i(t_i), \vartheta_{-i}(t_{-i})) \in c_i^{\vartheta(t)}(O_i^{\mu^f}(\alpha_{-i}^{\mathrm{id}}(t_{-i})))$ establish (*i*) of ex-post consistency of $(f(\Theta_i, \theta_{-i}))_{i\in N,\ \theta_{-i}\in\Theta_{-i}}$. For (*ii*) of ex-post consistency, for any deception/strategy profile α with $f \circ \vartheta \circ \alpha \neq f$, $\alpha^{\mathrm{id}} \circ \alpha = \alpha$ cannot be an EBE of μ^f because otherwise $g^f \circ \alpha^{\mathrm{id}} \circ \alpha = f \circ \vartheta \circ \alpha$ and hence by (*ii*) of ex-post implementation SCF $f \circ \vartheta \circ \alpha$ equals f, a contradiction. Thus, there is $i^* \in N$, $t^* \in T$ with $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(f(\Theta_{i^*}, \alpha_{-i^*}(t^*_{-i^*})))$ since $O_{i^*}^{\mu^f}((\alpha_j^{\mathrm{id}}(\alpha_j(t_j^*))_{j\neq i^*}) = f(\Theta_{i^*}, \alpha_{-i^*}(t^*_{-i^*}))$.

For the *sufficiency* of (*i*) of the theorem: By hypothesis, $(f(\Theta_i, \theta_{-i}))_{i\in N,\theta_{-i}\in\Theta_{-i}}$ is ex-post consistent with $F = \{f\}$. Then, α^{id} is a truthful EBE and $g^f \circ \alpha^{\mathrm{id}} = f \circ \vartheta$ thanks to quasi-ex-post incentive compatibility (implied by ex-post consistency). Further, if α^*

is an EBE, then $g^f \circ \alpha^* = f \circ \vartheta$; otherwise, $g^f(\alpha^*(t)) \neq f(\vartheta(t))$ for some $t \in T$. So, $f(\vartheta(\alpha^*(t))) \neq f(\vartheta(t))$ implies, by (ii) of ex-post consistency of $(f(\Theta_i, \theta_{-i}))_{i\in N,\ \theta_{-i}\in\Theta_{-i}}$, there is $i^* \in N$ and $t^* \in T$ with $g^f(\alpha^*(t^*)) = f(\vartheta(\alpha^*(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(f(\Theta_{i^*}, \alpha^*_{-i^*}(t^*_{-i^*})))$ as $g^f = f \circ \vartheta$, contradicting to α^* being an EBE as $O_{i^*}^{\mu^f}(\alpha^*_{-i^*}(t^*_{-i^*})) = f(\Theta_{i^*}, \alpha^*_{-i^*}(t^*_{-i^*}))$.

For (ii) of the theorem: The sufficiency holds trivially. On the other hand, if f is full-range and ex-post behavioral implementable by a general mechanism μ and σ is an EBE of μ with $g \circ \sigma = f \circ \vartheta$, then for all $i \in N$, $O_i^\mu(\sigma_{-i}(t_{-i})) = X$ for all $t_{-i} \in T_{-i}$ as $\vartheta_{-i}(T_{-i}) = \times_{j\neq i}\vartheta_j(T_j) = \Theta_{-i}$ (since $\vartheta_j : T_j \to \Theta_j$ is surjective for all $j \in N$), and for all $x \in X$, all $i \in N$, and all $\theta_{-i} \in \Theta_{-i}$, there is $\theta_i \in \Theta_i$ with $f(\theta) = x$ (because f is full-range). Hence, $\mathbb{S} = (S_i(\theta_{-i}))_{i\in N,\ \theta_{-i}\in\Theta_i}$ such that $S_i(\theta_{-i}) = X$ for all $i \in N$ and $\theta_{-i} \in \Theta_{-i}$ is ex-post consistent with $F = \{f\}$ by Theorem 1. Therefore, (i) of ex-post consistency implies for all $t \in T$, $f(\vartheta_i(t_i), \vartheta_{-i}(t_{-i})) \in c_i^{(\vartheta_i(t_i),\vartheta_{-i}(t_{-i}))}(X)$; (ii) of consistency implies for any α with $f \circ \vartheta \circ \alpha \neq f \circ \vartheta$, there is $i^* \in N$ and t^* such that $f(\vartheta(\alpha(t^*))) \notin c_{i^*}^{\vartheta(t^*)}(X)$. Now, consider SCF f's direct mechanism, μ^f, and observe that for all $i \in N$ and $O_i^{\mu^f}(t_{-i}) = f(\Theta_i, \vartheta_{-i}(t_{-i})) = X$ for all $t_{-i} \in T_{-i}$. Therefore, the truthtelling strategy profile α^{id} is an EBE of μ^f because (by (i) of ex-post consistency) for any $i \in N$ and $t_i \in T_i$, $g^f(\alpha_i^{\mathrm{id}}(t_i), \alpha_{-i}^{\mathrm{id}}(\tilde{t}_{-i})) = f(\vartheta_i(t_i), \vartheta(\tilde{t}_{-i})) \in c_i^{(\vartheta_i(t_i),\vartheta(\tilde{t}_{-i}))}(X)$ for all $\tilde{t}_{-i} \in T_{-i}$. For any other EBE $\tilde{\alpha}$ of direct mechanism μ^f and for any $\tilde{t} \in T$, it must be that $g^f(\tilde{\alpha}(\tilde{t})) = f(\vartheta(\tilde{t}))$. Otherwise, as $g^f(\tilde{\alpha}(\tilde{t})) = f(\hat{\theta})$ for some $\hat{\theta} \in \Theta$ due to the full-range condition, if $f(\hat{\theta}) \neq f(\vartheta(\tilde{t}))$, then (as $g^f = f \circ \vartheta$) we have $f(\vartheta(\tilde{\alpha}(\tilde{t}))) = f(\hat{\theta}) \neq f(\vartheta(\tilde{t}))$; so, by (ii) of ex-post consistency, there are $i^* \in N$ and $t^* \in T$ such that $g^f(\tilde{\alpha}(\theta^*)) = f(\vartheta(\tilde{\alpha}(\theta^*))) \notin c_{i^*}^{\vartheta(t^*)}(X)$, contradicting $\tilde{\alpha}$ being an EBE of the direct mechanism μ^f. ■

Finally, we note that Example 1 displays the use of part (i) of Theorem 3: Table 3 shows that the profile $(f(\Theta_i, \theta_{-i}))_{i\in N,\ \theta_{-i}\in\Theta_{-i}}$ is ex-post consistent with $F = \{\langle xzzy\rangle\}$.

C The Warning of de Clippel (2023)

We discuss situations in which a contradiction along the lines of de Clippel (2023) may emerge in our behavioral setting. To that regard, we construct an example mimicking the construction in the proof of de Clippel (2023, Proposition 1): Suppose that individuals' ex-post choices are singleton valued while the IIA does not hold for some individual's ex-post choices. Hence, there is an individual i, a choice-state $\theta \in \Theta$, a non-empty set of alternatives $T \in \mathcal{X}$, and an alternative $x \in T \setminus c_i^\theta(T)$ such that $c_i^\theta(T) \neq c_i^\theta(T \setminus \{x\})$. Given i and her type t_i with $\vartheta_i(t_i) = \theta_i$, if there are only two distinct profiles of others, $t_{-i}, \tilde{t}_{-i} \in T_{-i}$ with $\vartheta_{-i}(t_{-i}) = \theta_{-i} \neq \vartheta_{-i}(\tilde{t}_{-i})$ and hence $\vartheta(t) = \theta \neq (\vartheta_i(t_i), \vartheta_{-i}(\tilde{t}_{-i}))$,

then we can construct the following set of acts: $\tilde{\mathbf{A}}_i := \{\mathbf{a}_i, \mathbf{a}'_i, \mathbf{a}''_i, \mathbf{a}'''_i\} \bigcup \left(\cup_{y \in Y,\ \tilde{y} \in \tilde{Y}} \{\mathbf{a}_i^{y,\tilde{y}}\}\right)$ where these acts are as specified in Table 15 and $Y, \tilde{Y} \in \mathcal{X}$ are as follows:

	t_{-i}	$\tilde{t}_{-i}$
$\mathbf{a}_i$	$c_i^{\vartheta(t)}(T)$	$c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(T \setminus \{x\})$
$\mathbf{a}'_i$	$c_i^{\vartheta(t)}(T \setminus \{c_i^{\vartheta(t_i)}(T)\})$	$c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(T \setminus \{x\})$
$\mathbf{a}''_i$	$c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(T \setminus \{x\})$	$c_i^{\vartheta(t)}(T)$
$\mathbf{a}'''_i$	x	$c_i^{\vartheta(t)}(T)$
$\mathbf{a}_i^{y,\tilde{y}}$	y	$\tilde{y}$

Table 15: An example in conjunction with Property STP*.

$$Y = T \setminus \left\{x, c_i^{\vartheta(t)}(T), c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(T \setminus \{x\}), c_i^{\vartheta(t)}(T \setminus \{c_i^{\vartheta(t)}(T)\})\right\},$$
$$\tilde{Y} = T \setminus \left\{x, c_i^{\vartheta(t)}(T), c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(T \setminus \{x\})\right\},$$

We let $\mathbf{A}_i^* := \tilde{\mathbf{A}}_i \setminus \{\mathbf{a}_i\}$. Then, we observe that $\tilde{\mathbf{A}}_i(t_{-i}) = T$, $\tilde{\mathbf{A}}_i(\tilde{t}_{-i}) = T \setminus \{x\}$, $\mathbf{A}_i^*(t_{-i}) = T \setminus \{c_i^{\vartheta(t)}(T)\}$, and $\mathbf{A}_i^*(\tilde{t}_{-i}) = T \setminus \{x\}$. Thus, by Property STP*, $\mathbf{a}_i \in \mathbf{C}_i^{t_i}(\tilde{\mathbf{A}}_i)$ as $\mathbf{a}_i(t_{-i}) = c_i^{\vartheta(t)}(\tilde{\mathbf{A}}_i(t_{-i}))$ and $\mathbf{a}_i(\tilde{t}_{-i}) = c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(\tilde{\mathbf{A}}_i(\tilde{t}_{-i}))$. Similarly, $\mathbf{a}'_i \in \mathbf{C}_i^{t_i}(\mathbf{A}_i^*)$ since $\mathbf{a}'_i(t_{-i}) = c_i^{\vartheta(t)}(\mathbf{A}_i^*(t_{-i}))$ and $\mathbf{a}'_i(\tilde{t}_{-i}) = c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(\mathbf{A}_i^*(\tilde{t}_{-i}))$.

We need the following additional requirements to reach a contradiction as in de Clippel (2023): Individual i of type t_i should perceive acts $\mathbf{a}_i$ and $\mathbf{a}''_i$ to be equivalent to each other on grounds of $\mathbf{a}_i(t_{-i}) = \mathbf{a}''_i(\tilde{t}_{-i})$ and $\mathbf{a}_i(\tilde{t}_{-i}) = \mathbf{a}''_i(t_{-i})$. That is, when considering t_{-i} and $\tilde{t}_{-i}$, only the underlying alternatives associated with these acts matter to her. As a result, she perceives the act that delivers x' at t_{-i} and y' at $\tilde{t}_{-i}$ to be equivalent to another that provides y' at t_{-i} and x' at $\tilde{t}_{-i}$ where $x', y' \in X$. For example, this happens under probabilistic sophistication when i's belief when her type is t_i is such that t_{-i} and $\tilde{t}_{-i}$ are equally likely and i of type t_i evaluates acts by the lotteries they induce.

Indeed, at the heart of this contradiction lies the individual perceiving two different states as equivalent. We model the equivalence perception of individual i of type t_i over the set of all others' types via an equivalence relation $\doteqdot$ defined on T_{-i} and let the equivalence class of $\bar{t}_{-i}$ be $\mathcal{P}_{i,t_i}(\bar{t}_{-i}) := \{t'_{-i} \in T_{-i} \mid t'_{-i} \doteqdot \bar{t}_{-i}\}$.[28] For any i of type t_i, the relation $\doteqdot$ partitions any set of states $\bar{T}_{-i} \subset T_{-i}$ into equivalence classes.

As a result of this perception equivalence, individual i of type t_i perceives two acts $\mathbf{a}_i^{(1)}$ and $\mathbf{a}_i^{(2)}$ as equivalent whenever for any $t'_{-i}, t''_{-i} \in T_{-i}$ with $t''_{-i} \in \mathcal{P}_{i,t_i}(t'_{-i})$, $\mathbf{a}_i^{(1)}(t'_{-i}) = \mathbf{a}_i^{(2)}(t''_{-i})$, $\mathbf{a}_i^{(1)}(t''_{-i}) = \mathbf{a}_i^{(2)}(t'_{-i})$, and $\mathbf{a}_i^{(1)}(t'''_{-i}) = \mathbf{a}_i^{(2)}(t'''_{-i})$ for all $t'''_{-i} \in T_{-i} \setminus \{t'_{-i}, t''_{-i}\}$; we

[28] An equivalence relation is a binary relation that is reflexive, symmetric, and transitive.

denote such a situation by $\mathbf{a}_i^{(1)}\mathcal{I}_{i,t_i}\mathbf{a}_i^{(2)}$. We also need to assume that the perception equivalence relation of individual i of type t_i satisfies the following: If $\mathbf{a}_i^{(1)}\mathcal{I}_{i,t_i}\mathbf{a}_i^{(2)}$, then $\mathbf{a}_i^{(1)} \in \mathbf{C}_i^{t_i}(\mathbf{A}_i')$ if and only if $\mathbf{a}_i^{(2)} \in \mathbf{C}_i^{t_i}(\mathbf{A}_i')$ for all $\mathbf{A}_i' \subset \mathbf{A}_i$ with $\mathbf{a}_i^{(1)}, \mathbf{a}_i^{(2)} \in \mathbf{A}_i'$.

For any i of type t_i, the relation $\mathcal{I}_{i,t_i}$ partitions any $\mathbf{A}_i' \subset \mathbf{A}_i$ into equivalence classes. We assume that i of type t_i perceives two sets of acts $\mathbf{A}_i'$ and $\mathbf{A}_i''$ as equivalent if the collection of equivalence classes in $\mathbf{A}_i$ that the acts in $\mathbf{A}_i'$ and $\mathbf{A}_i''$ belong to are equal to one another. With a slight abuse of notation, we denote such a case by $\mathbf{A}_i'\mathcal{I}_{i,t_i}\mathbf{A}_i''$. Formally, $\mathbf{A}_i^{(1)}\mathcal{I}_{i,t_i}\mathbf{A}_i^{(2)}$ if for all $\bar{\mathbf{a}}_i \in \mathbf{A}_i^{(k)}$, $\mathcal{I}_{i,t_i}(\bar{\mathbf{a}}_i) \cap \mathbf{A}_i^{(\ell)} \neq \emptyset$ for all $k, \ell = 1, 2$ where $\mathcal{I}_{i,t_i}(\bar{\mathbf{a}}_i)$ is the equivalence class of $\bar{\mathbf{a}}_i$ with respect to $\mathcal{I}_{i,t_i}$.

Moreover, we assume that interim choices of i of type t_i from a set of acts respect the resulting equivalence classes so that the interim choices are singleton-valued up to equivalence classes with respect to $\mathcal{I}_{i,t_i}$: For any two sets of acts $\mathbf{A}_i^{(1)}$ and $\mathbf{A}_i^{(2)}$ with $\mathbf{A}_i^{(1)}\mathcal{I}_{i,t_i}\mathbf{A}_i^{(2)}$, $\mathbf{a}_i^{(1)} \in \mathbf{C}_i^{t_i}(\mathbf{A}_i^{(1)})$ and $\mathbf{a}_i^{(2)} \in \mathbf{C}_i^{t_i}(\mathbf{A}_i^{(2)})$ implies $\mathbf{a}_i^{(1)}\mathcal{I}_{i,t_i}\mathbf{a}_i^{(2)}$.

Finally, going back to our example, we have $\mathbf{a}_i\mathcal{I}_{i,t_i}\mathbf{a}_i''$ as $\mathbf{a}_i = \langle c_i^{\vartheta(t)}(T),\ c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(T \setminus \{x\})\rangle$ and $\mathbf{a}_i'' = \langle c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(T \setminus \{x\}),\ c_i^{\vartheta(t)}(T)\rangle$. Further, $\tilde{\mathbf{A}}_i\mathcal{I}_{i,t_i}\mathbf{A}_i^*$ because $\mathbf{A}_i^* = \tilde{\mathbf{A}}_i \setminus \{\mathbf{a}_i\}$ and $\mathbf{a}_i\mathcal{I}_{i,t_i}\mathbf{a}_i''$ and $\mathbf{a}_i'' \in \mathbf{A}_i^* \subset \tilde{\mathbf{A}}_i$. Recall that Property STP* implies $\mathbf{a}_i \in \mathbf{C}_i^{t_i}(\tilde{\mathbf{A}}_i)$ and $\mathbf{a}_i' \in \mathbf{C}_i^{t_i}(\mathbf{A}_i^*)$. The desired contradiction emerges as $\mathbf{a}_i\mathcal{I}_{i,t_i}\mathbf{a}_i'$ does not hold since $\mathbf{a}_i(t_{-i}) = c_i^{\vartheta(t)}(T) \neq c_i^{\vartheta(t)}(T \setminus c_i^{\vartheta(t)}(T)) = \mathbf{a}_i'(t_{-i})$ but $\mathbf{a}_i(\tilde{t}_{-i}) = \mathbf{a}_i'(\tilde{t}_{-i}) = c_i^{(\vartheta_i(t_i),\vartheta_{-i}(\tilde{t}_{-i}))}(T \setminus \{x\})$.

D Robustness Properties

To analyze the robustness properties of ex-post behavioral implementation, we provide the following details regarding individuals' information and type spaces: Let $\mathcal{T}$ be *the set of all permissible type spaces* of the form $(T_i, \vartheta_i)_{i\in N}$ where the association of individual i's type (including her beliefs about others' types) with her ex-post choice-type is captured by $\vartheta_i : T_i \to \Theta_i$, a surjection, for all $i \in N$. Thus, for all $(T_i, \vartheta_i)_{i\in N}, (\tilde{T}_i, \tilde{\vartheta}_i)_{i\in N} \in \mathcal{T}$, we have $\vartheta_i(T_i) = \tilde{\vartheta}_i(\tilde{T}_i) = \Theta_i$ for all $i \in N$. With a slight abuse of notation, we denote a *type space* by $\mathcal{T} = (T, \vartheta)$ where $T = \times_{i\in N} T_i$ and $\vartheta : T \to \Theta$ is defined as $\vartheta(t) = \times_{i\in N}\vartheta_i(t_i)$ for all $t \in T$.[29] We assume that the realized type space is commonly observed by the individuals but not the planner. Each individual observes only her own type but not that of the others. We formalize *individual i's private information* at the interim stage via

[29]This formulation parallels that of Bergemann and Morris (2011). Indeed, in the rational domain, a type space is denoted by a triplet $\mathcal{T} = (T_i, \vartheta_i, \pi_i)_{i\in N}$ where $\pi_i : T_i \to \Delta(T_{-i})$ denotes the probabilistic beliefs of individuals over other individuals' types. We do not include such probabilistic beliefs in our type space formulation because in our behavioral setup, individuals' assessments of others' types are entangled in their interim choices. Thus, it is not clear how to obtain these assessments via probabilistic beliefs when individuals are not necessarily rational.

$\Omega_i := \{((T,\vartheta),t_i) \mid (T,\vartheta) \in \mathcal{T}, t_i \in T_i\}$; so, the private information of individual i at the interim stage, $\omega_i = ((T,\vartheta),t_i) \in \Omega_i$, represents the situation in which individual i observes the realized type space $(T,\vartheta) \in \mathcal{T}$ along with her type $t_i \in T_i$.

Given a permissible type space $\mathcal{T} = (T,\vartheta) \in \mathcal{T}$ and individual i of type $t_i \in T_i$, $\mathbf{A}_i^{\mathcal{T}}$ denotes the set of all acts that i of type t_i faces at $\mathcal{T}$ (i.e., $\mathbf{A}_i^{\mathcal{T}} := \{\mathbf{a}_i \mid \mathbf{a}_i : T_{-i} \to X\}$).

In this framework, a strategy of individual i in a mechanism $\mu = (M,g)$ is $\sigma_i : \Omega_i \to M_i$. Let $\sigma_i = (\sigma_i^{\mathcal{T}})_{\mathcal{T}\in\mathcal{T}}$ where $\sigma_i^{\mathcal{T}}$ is the projection of σ_i on the permissible type space $\mathcal{T} = (T,\vartheta)$. We refer to the set of acts individual i can unilaterally generate at the commonly observed type space $\mathcal{T}$ when the other individuals use $\sigma_{-i}^{\mathcal{T}} := (\sigma_j^{\mathcal{T}})_{j\neq i}$ as *individual i's opportunity set of acts at $\mathcal{T}$ under μ for $\sigma_{-i}^{\mathcal{T}}$*. Formally, for any $i \in N$, $\mathbf{O}_i^{\mu}(\sigma_{-i}) := (\mathbf{O}_{i,\mathcal{T}}^{\mu}(\sigma_{-i}^{\mathcal{T}}))_{\mathcal{T}\in\mathcal{T}}$ where

$$\mathbf{O}_{i,\mathcal{T}}^{\mu}(\sigma_{-i}^{\mathcal{T}}) := \left\{\mathbf{a}_i^{\mathcal{T}} \in \mathbf{A}_i^{\mathcal{T}} \mid \exists m_i \in M_i \text{ s.t. } \mathbf{a}_i^{\mathcal{T}}(t_{-i}) = g(m_i, \sigma_{-i}^{\mathcal{T}}(t_{-i})) \text{ for all } t_{-i} \in T_{-i}\right\}.$$

A strategy profile $\sigma^{\mathcal{T}} = (\sigma_i^{\mathcal{T}})_{i\in N}$ with $\sigma_i^{\mathcal{T}} : T_i \to M_i$ is a BIE of μ at type space $\mathcal{T} = (T,\vartheta)$ if for all $i \in N$ and $t_i \in T_i$, $[g \circ \sigma^{\mathcal{T}}]_{(i,t_i)} \in \mathbf{C}_i^{t_i}(\mathbf{O}_{i,\mathcal{T}}^{\mu}(\sigma_{-i}^{\mathcal{T}}))$ where $[g \circ \sigma^{\mathcal{T}}]_{(i,t_i)} : T_{-i} \to X$ is the interim act i of type t_i faces at type space $\mathcal{T}$ defined by $[g \circ \sigma^{\mathcal{T}}]_{(i,t_i)}(t_{-i}) = g(\sigma_i^{\mathcal{T}}(t_i), \sigma_{-i}^{\mathcal{T}}(t_{-i}))$ for all $t_{-i} \in T_{-i}$. A strategy profile $\sigma^{\mathcal{T}} = (\sigma_i^{\mathcal{T}})_{i\in N}$ is an EBE of μ at $\mathcal{T} = (T,\vartheta)$ if for all $i \in N$ and $t_i \in T_i$, $g(\sigma_i^{\mathcal{T}}(t_i), \sigma_{-i}^{\mathcal{T}}(t_{-i})) \in c_i^{(\vartheta_i(t_i),\vartheta_{-i}(t_{-i}))}(\mathbf{O}_{i,\mathcal{T}}^{\mu}(\sigma_{-i}^{\mathcal{T}})(t_{-i}))$ for all $t_{-i} \in T_{-i}$ where $\mathbf{O}_{i,\mathcal{T}}^{\mu}(\sigma_{-i}^{\mathcal{T}})(t_{-i}) := O_i^{\mu}(\sigma_{-i}^{\mathcal{T}}(t_{-i}))$.

The following result implies that EBE is robust in the following sense:

Proposition 10. *An EBE of mechanism μ at a permissible type space induces an outcome equivalent EBE of μ at every other permissible type space.*

Proof. Let a strategy profile $\sigma^{\mathcal{T}}$ given by $\sigma_j^{\mathcal{T}} : T_j \to M_j$ for all $j \in N$ be an EBE of μ at $\mathcal{T} = (T,\vartheta) \in \mathcal{T}$. Then, for all $i \in N$ and all $t_i \in T_i$, $g(\sigma_i^{\mathcal{T}}(t_i), \sigma_{-i}^{\mathcal{T}}(t_{-i})) \in c_i^{(\vartheta_i(t_i),\vartheta_{-i}(t_{-i}))}(\mathbf{O}_{i,\mathcal{T}}^{\mu}(\sigma_{-i}^{\mathcal{T}})(t_{-i}))$ for all $t_{-i} \in T_{-i}$. For any other permissible type space $\tilde{\mathcal{T}} = (\tilde{T},\tilde{\vartheta})$, define $\sigma^{\tilde{\mathcal{T}}}$ as follows: For all $j \in N$ and $\tilde{t}_j \in \tilde{T}_j$, let $\sigma_j^{\tilde{\mathcal{T}}}(\tilde{t}_j) = \sigma_j^{\mathcal{T}}(t_j)$ for some $t_j^* \in T_j$ with $\vartheta_j(t_j^*) = \tilde{\vartheta}_j(\tilde{t}_j)$. Because $\vartheta_j : T_j \to \Theta_j$ and $\tilde{\vartheta}_j : \tilde{T}_j \to \Theta_j$ are surjective for all $j \in N$, $\sigma^{\tilde{\mathcal{T}}}$ is a well-defined strategy profile at $\tilde{\mathcal{T}}$. Then, for all $i \in N$ and all $\tilde{t}_i \in \tilde{T}_i$ together with its associated type $t_i^* \in T_i$ such that $\vartheta_i(t_i^*) = \tilde{\vartheta}_i(\tilde{t}_i)$, we observe that $g(\sigma_i^{\mathcal{T}}(t_i^*), \sigma_{-i}^{\mathcal{T}}(t_{-i}^*)) = g(\sigma_i^{\tilde{\mathcal{T}}}(\tilde{t}_i), \sigma_{-i}^{\tilde{\mathcal{T}}}(\tilde{t}_{-i}))$ and $\mathbf{O}_{i,\mathcal{T}}^{\mu}(\sigma_{-i}^{\mathcal{T}}))(t_{-i}^*) = \mathbf{O}_{i,\tilde{\mathcal{T}}}^{\mu}(\sigma^{\tilde{\mathcal{T}}_{-i}}))(\tilde{t}_{-i})$. Therefore, for all $i \in N$ and all $\tilde{t}_i \in \tilde{T}_i$, $g(\sigma_i^{\tilde{\mathcal{T}}}(\tilde{t}_i), \sigma_{-i}^{\tilde{\mathcal{T}}}(\tilde{t}_{-i})) \in c_i^{(\tilde{\vartheta}_i(\tilde{t}_i),\tilde{\vartheta}_{-i}(\tilde{t}_{-i}))}(\mathbf{O}_{i,\tilde{\mathcal{T}}}^{\mu}(\sigma_{-i}^{\tilde{\mathcal{T}}}))(\tilde{t}_{-i})$ for all $\tilde{t}_{-i} \in \tilde{T}_{-i}$, establishing that $\sigma^{\tilde{\mathcal{T}}}$ is an outcome equivalent EBE of μ at $\tilde{\mathcal{T}}$. ■

Proposition 10 implies that ex-post behavioral implementability of an SCS F is independent of the permissible type space. Consequently, one can dismiss concerns about

individuals' beliefs (interim assessment of others' types) not affecting their ex-post choices and focus on a type space that is in one-to-one correspondence with the (ex-post) choice-states. Thus, the following type space suffices when analyzing ex-post behavioral implementability: $\mathcal{T}^{\Theta} = (\Theta, \vartheta^{\Theta})$ where for any $i \in N$, $\vartheta_i^{\Theta} : \Theta_i \to \Theta_i$ is the identity function.

These imply that every EBE of mechanism μ at a permissible type space induces an outcome equivalent BIE of μ at every other permissible type space under Property STP*.

Acknowledgements

We are grateful to Tilman Börgers, Alessandro Pavan, and the anonymous reviewers for their valuable comments. We used large language models, including ChatGPT and Gemini, solely for language editing. Any remaining errors are ours.

References

Barlo, M., & Dalkiran, N. A. (2009). Epsilon-Nash implementation. *Economics Letters*, *102*(1), 36–38.

Barlo, M., & Dalkıran, N. A. (2022). Computational implementation. *Review of Economic Design*, *26*(4), 605–633.

Barlo, M., & Dalkıran, N. A. (2023a). Behavioral implementation under incomplete information. *Journal of Economic Theory*, 105738.

Barlo, M., & Dalkıran, N. A. (2023b). Implementation with missing data. *Mimeo*.

Barlo, M., & Dalkıran, N. A. (2026). Robust behavioral implementation. *Mimeo*.

Bergemann, D., & Morris, S. (2005). Robust mechanism design. *Econometrica*, *73*(6), 1771–1813.

Bergemann, D., & Morris, S. (2008). Ex post implementation. *Games and Economic Behavior*, *63*(2), 527–566.

Bergemann, D., & Morris, S. (2009). Robust implementation in direct mechanisms. *The Review of Economic Studies*, *76*(4), 1175–1204.

Bergemann, D., & Morris, S. (2011). Robust implementation in general mechanisms. *Games and Economic Behavior*, *71*(2), 261–281.

Bergemann, D., & Morris, S. (2017). Belief-free rationalizability and informational robustness. *Games and Economic Behavior*, *104*, 744–759.

Chen, Y.-C., Holden, R., Kunimoto, T., Sun, Y., & Wilkening, T. (2023). Getting dynamic implementation to work. *Journal of Political Economy*, *131*(2), 285–387.

Chen, Y.-C., Kunimoto, T., Sun, Y., & Xiong, S. (2021). Rationalizable implementation in finite mechanisms. *Games and Economic Behavior*, *129*, 181–197.

Chen, Y.-C., Mueller-Frank, M., & Pai, M. M. (2022). Continuous implementation with direct revelation mechanisms. *Journal of Economic Theory*, *201*, 105422.

Dean, M., Kıbrıs, Ö., & Masatlioglu, Y. (2017). Limited attention and status quo bias. *Journal of Economic Theory*, *169*, 93–127.

de Clippel, G. (2014). Behavioral implementation. *American Economic Review*, *104*(10), 2975–3002.

de Clippel, G. (2023). Departures from preference maximization and violations of the sure-thing principle. *Mimeo*.

de Clippel, G., & Eliaz, K. (2012). Reason-based choice: A bargaining rationale for the attraction and compromise effects. *Theoretical Economics*, *7*(1), 125–162.

de Clippel, G., Saran, R., & Serrano, R. (2019). Level-mechanism design. *The Review of Economic Studies*, *86*(3), 1207–1227.

Eliaz, K. (2002). Fault tolerant implementation. *The Review of Economic Studies*, *69*(3), 589–610.

Hagiwara, M. (2025). Behavioral subgame perfect implementation. *Journal of Economic Behavior & Organization*, *233*, 106992.

Hayashi, T., Jain, R., Korpela, V., & Lombardi, M. (2023). Behavioral strong implementation. *Economic Theory*, 1–31.

Holmström, B., & Myerson, R. B. (1983). Efficient and durable decision rules with incomplete information. *Econometrica*, 1799–1819.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, *9*(1), 90–98.

Hurwicz, L. (1986). On the implementation of social choice rules in irrational societies. *Social Choice and Public Decision Making: Essays in Honor of Kenneth J. Arrow*.

Jackson, M. O. (1991). Bayesian implementation. *Econometrica*, 461–477.

Jain, R., Korpela, V., & Lombardi, M. (2025). Two-player rationalizable implementation. *Journal of Economic Theory*, 106031.

Jain, R., & Lombardi, M. (2022). On interim rationalizable monotonicity. *Available at SSRN 4106795*.

Jain, R., Lombardi, M., & Müller, C. (2023). An alternative equivalent formulation for robust implementation. *Games and Economic Behavior*.

Korpela, V. (2012). Implementation without rationality assumptions. *Theory and Decision*, *72*(2), 189–203.

Korpela, V., & Lombardi, M. (2019). (Interim) Bayesian efficiency implies two-agent Bayesian implementation. *Available at SSRN 3134454*.

Kucuksenel, S. (2012). Behavioral mechanism design. *Journal of Public Economic Theory*, *14*(5), 767–789.

Kunimoto, T., & Saran, R. (2024). Robust implementation in rationalizable strategies in general mechanisms. *Mimeo*.

Kunimoto, T., Saran, R., & Serrano, R. (2023). Interim rationalizable implementation of functions. *Mathematics of Operations Research*.

Kunimoto, T., Saran, R., & Serrano, R. (2025). Rationalizable incentives: Interim rationalizable implementation of correspondences.

Manzini, P., & Mariotti, M. (2007). Sequentially rationalizable choice. *American Economic Review*, *97*(5), 1824–1839.

Masatlioglu, Y., & Ok, E. A. (2014). A Canonical Model of Choice with Initial Endowments. *The Review of Economic Studies*, *81*(2), 851-883.

Maskin, E. (1999). Nash equilibrium and welfare optimality. *The Review of Economic Studies*, *66*(1), 23–38.

Moore, J., & Repullo, R. (1990). Nash implementation: a full characterization. *Econometrica*, 1083–1099.

Ohashi, Y. (2012). Two-person ex post implementation. *Games and Economic Behavior*, *75*(1), 435–440.

Ok, E. A., Ortoleva, P., & Riella, G. (2015). Revealed (p)reference theory. *American Economic Review*, *105*(1), 299–321.

Ollár, M., & Penta, A. (2017). Full implementation and belief restrictions. *American Economic Review*, *107*(8), 2243–2277.

Palfrey, T. R., & Srivastava, S. (1987). On Bayesian implementable allocations. *The Review of Economic Studies*, *54*(2), 193–208.

Penta, A. (2015). Robust dynamic implementation. *Journal of Economic Theory*, *160*, 280–316.

Postlewaite, A., & Schmeidler, D. (1986). Implementation in differential information economies. *Journal of Economic Theory*, *39*(1), 14–33.

Rubbini, G. (2023). Mechanism design without rational expectations. *arXiv preprint arXiv:2305.07472*.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59.

Saran, R. (2011). Menu-dependent preferences and revelation principle. *Journal of Economic Theory*, *146*(4), 1712–1720.

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association*, *46*(253), 55–67.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Sen, A. K. (1971). Choice functions and revealed preference. *The Review of Economic Studies*, *38*(3), 307–317.

Xiong, S. (2023). Rationalizable implementation of social choice functions: complete characterization. *Theoretical Economics*, *18*, 197–230.