Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1918

Incentive compatibility and belief restrictions

Mariann Ollár and Antonio Penta

September 2025

Incentive Compatibility and Belief Restrictions*

Mariann Ollár
NYU Shanghai and TSE

Antonio Penta
ICREA, UPF, BSE and TSE

Abstract

We study a framework for robust mechanism design that can accommodate various degrees of robustness with respect to agents' beliefs, which encompasses both the belief-free and Bayesian settings as special cases. For general *belief restrictions*, we characterize the set of incentive compatible direct mechanisms in general environments with interdependent values. Our main results, which we obtain based on a *first-order approach*, inform the design of transfers via 'belief-based' terms to attain incentive compatibility. In environments that satisfy a property of *generalized independence*, our results imply a *robust* version of *revenue equivalence* in non-Bayesian settings. Instead, under a notion of *comovement* between types and beliefs, which extends the idea of correlated information to non-Bayesian settings, we show that any allocation rule can be implemented, even if standard single-crossing and monotonicity conditions do not hold. Yet, unless the environment is Bayesian, information rents typically remain, and they decrease monotonically as the robustness requirements are weakened.

Keywords: Moment Conditions, Robust Mechanism Design, Incentive Compatibility, Interdependent Values, Belief Restrictions

JEL: D62, D82, D83

1 Introduction

Mechanism design has greatly succeeded in deepening our understanding of incentives under private information, and it has had a dramatic impact on the design of real world mechanisms and institutions. Yet, the classical approach also features some important limitations, particularly due to the strong assumptions on agents' beliefs that are implicit in standard models, and the role they play in several results. The 'Full Surplus Extraction' results of Crémer and McLean (1985, 1988) are notorious examples of findings that cast doubt on the

*We thank Thomas Mariotti, Ludvig Sinander, Jianrong Tian, Juuso Toikka and Gabor Virag for their comments and suggestions. We also thank the audiences at the Workshop on the Design of Strategic Interaction (Venice, 2023), the Inaugural Janeway Institute Microeconomic Theory Conference (Cambridge, 2024), Workshop on Contracts, Incentives and Information (CEPR and CCA, Turin, 2024), the Conference on Mechanism and Institution Design (Budapest, 2024), the Durham Economic Theory Conference (2024), the Lancaster Game Theory Conference (2023), the CUHK Workshop on Economic Theory (2023), the Bonn Christmas Microeconomic Theory Conference (2024), the Makris Symposiium in Economic Theory (2025) and at seminars at SMU, NUS, NYU, HKUST, UPF, Northwestern, NYU-Shanghai, Manchester. The BSE acknowledges the financial support of the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S). Antonio Penta acknowledges the financial support of the ERC Consolidator Grant n. 101089139.

adequacy of the classical mechanism design paradigm.[1] Together with several other results in the literature, they motivated Wilson (1987)'s famous call for a "repeated weakening of common knowledge assumptions". In response, a *robust* approach to mechanism design developed, where mechanisms are required to perform well for a large set of beliefs.[2]

In this paper we provide a systematic analysis of robust mechanism design under general belief restrictions. We characterize the set of incentive compatible mechanisms, we identify a novel and tractable design principle, and discuss several of its implications. These include a *robust* version of *revenue equivalence* for non-Bayesian settings, as well as constructive implementation results for environments that violate standard single-crossing and monotonicity conditions. We also provide a notion of *comovement* between types and beliefs that extends the idea of correlation to non-Bayesian settings, and we show that it enables permissive implementation results, while at the same time avoiding the pitfalls of the classical Bayesian paradigm that we discussed above. Thus, on the one hand we provide design suggestions and results that do not rely on the full distributional specificity of beliefs; on the other, we show that the latter are the ultimate responsible of the 'disturbing' FSE results.

More specifically, we model agents' beliefs as *belief restrictions*, $\mathcal{B} = ((B_{\theta_i})_{\theta_i \in \Theta_i})_{i \in I}$, where each type $\theta_i \in \Theta_i$ of an agent is endowed with a set of beliefs about others' types, $B_{\theta_i} \subseteq \Delta(\Theta_{-i})$, that the designer regards as possible. This way, we accommodate as special cases the classical Bayesian framework (where all such sets are singletons), the belief-free setting of Bergemann and Morris (2005) (where $B_{\theta_i} = \Delta(\Theta_{-i})$ for all i and $\theta_i \in \Theta_i$), but also the intermediate cases where the designer can rely on some, but not full, information about agents' beliefs. For instance, the designer may have information about agents' beliefs over some *moment* of the distributions of types, but not necessarily know the exact distribution. The framework also accommodates restrictions motivated by behavioral economics research, which cannot be cast within the standard paradigm, such as incomplete preferences *a la* Bewley (2002), or models with belief distortions (e.g., Gagnon-Bartsch et al. (2021) and Gagnon-Bartsch and Rosato (2024)).[3] Within these settings, we say that a direct mechanism is $\mathcal{B}$-*incentive compatible* ($\mathcal{B}$-IC) if truthful revelation is a mutual best-response, for all types and for all beliefs in the belief restrictions. Thus, depending on the belief restrictions, $\mathcal{B}$-IC generalizes both belief-free (or *ex-post*) and Bayesian (or *interim*) incentive compatibility, as well as intermediate notions of robustness.[4]

[1] For instance, Crémer and McLean (1988, p.1254): "Economic intuition and informal evidence (we know of no way to test such a proposition) suggest that this result is counterfactual, and several explanations can be suggested." The influential critique of Neeman (2004) may also be ascribed to this view.

[2] This approach was put forward by Bergemann and Morris (2005, 2009a,b), who studied partial, full, and virtual implementation in *belief-free* settings. The related literature is discussed in Section 6.

[3] The general framework to accommodate varying degrees of robustness was introduced by Ollár and Penta (2017) to study how beliefs can be used to attain *full implementation*, taking incentive compatibility as given. Here, in contrast, we tackle the more fundamental issue of incentive compatibility.

[4] In Section 2 we discuss several foundations for this notion, and the sense in which the restriction to direct mechanisms and $\mathcal{B}$-IC is without loss for our purposes (see Ennuschat and Penta, 2025). From a methodological perspective, we depart from the now dominant envelope approach and revisit instead the classical *first-order approach* (Rogerson, 1985; Jewitt, 1988). Carvajal and Ely (2013) also studied incentive

For the sake of illustration, first consider the problem of implementing an allocation rule, $d : \Theta \to X$, in a belief-free setting. As we show, this is possible if and only it is attained by the *canonical transfers*: namely, the transfers that are pinned down by the necessary first-order conditions for truthful revelation to be an ex-post equilibrium of the direct mechanism (e.g., if d is the efficient allocation rule, then the canonical transfers coincide with generalized VCG transfers). Under standard single-crossing conditions, the ex-post payoff functions that they induce are concave at each truthful profile if and only if the allocation rule is increasing, in which case truthful revelation is an ex-post equilibrium, and incentive compatibility is attained in a belief-free sense (ex-post incentive compatibility, ep-IC). But if either single-crossing or monotonicity fail, then the second-order conditions are not met, and ep-IC is not possible. In those cases, suitable modifications of the transfers may restore incentive compatibility, but only by relying on information about beliefs. Whether this is possible, or how, depends on the information that is available to the designer.

Given some belief restrictions, $\mathcal{B}$, suppose that a $\mathcal{B}$-IC transfer scheme can be obtained by an additive modification of the canonical transfers. Since, by construction, the canonical transfers ensure that truthful revelation satisfies the first-order conditions in the ex-post sense, so they do for all beliefs in $\mathcal{B}$. Hence, if an additive modification of the canonical transfers yields a $\mathcal{B}$-IC transfer scheme, then it must be that the added term also satisfies the first-order conditions, for all beliefs in the belief sets, while at the same time ensure that the payoff functions they induce have the right curvature.

Theorems 1 and 2 in Section 3 show that this intuition is fully general: for any belief-restrictions $\mathcal{B} = ((B_{\theta_i})_{\theta_i \in \Theta_i})_{i \in I}$, any $\mathcal{B}$-IC transfer can be written as

$$t_i(m) = t_i^*(m) + \beta_i(m),$$

where (letting $m \in M = \Theta$ denote a generic message profile in the direct mechanism) $t_i^* : M \to \mathbb{R}$ denotes the *canonical transfers*, which by construction only depend on the allocation rule and agents' preferences, and $\beta_i : M \to \mathbb{R}$ is a *belief-based term* such that, for all θ_i and $b_{\theta_i} \in B_{\theta_i}$, it must satisfy two conditions: (i) $\mathbb{E}^{b_{\theta_i}}\left[\frac{\partial \beta_i}{\partial m_i}(\theta_i, \theta_{-i})\right] = 0$, and (ii) $\mathbb{E}^{b_{\theta_i}}\left[\frac{\partial^2 \beta_i}{\partial^2 m_i}(\theta_i, \theta_{-i})\right] \leq -\mathbb{E}^{b_{\theta_i}}\left[\frac{\partial^2 U_i^*}{\partial^2 m_i}(\theta_i, \theta_{-i})\right]$ (where $U_i^*(\cdot)$ denotes the payoff function induced by the canonical transfers). Furthermore, a slight strengthening of the latter condition is also sufficient (Theorem 2). These results are further generalized by Theorem 3, which provides a tight characterization that highlights the role of belief-based terms in overcoming failures of standard single-crossing and monotonicity conditions.

These results formalize a general principle, according to which designing $\mathcal{B}$-IC transfer schemes boils down to designing belief-based terms that satisfy conditions (i) and (ii) above. The bite of these conditions depends on the richness of the belief sets, which determine the set of suitable belief-based terms and, hence, the set of incentive compatible transfers. For

compatibility in settings where the envelope approach cannot be used, albeit only in Bayesian settings.

instance, if the belief sets are constant across a player's types (what we call *generalized independence*), then the only β_i-terms that satisfy condition (i) do not affect agents' incentives to report truthfully, and hence $\mathcal{B}$-IC is possible if and only if it is attained by the canonical transfers (Corollary 1).[5] In these cases, the information about beliefs cannot be used to attain implementation when t^* does not. Outside of these settings, however, belief-based terms can greatly expand the possibility of implementation: As we show in Section 4, if types and belief sets 'comove' in a precise sense, then implementation can be attained even in settings that do not satisfy standard single-crossing and monotonicity conditions.

Specifically, in Section 4 we show that a weak *responsive moment condition* suffices to make *any* allocation rule $d : \Theta \to X$ incentive compatible, in any environment, via the suitable design of a belief-based term (Proposition 1). Loosely speaking, this condition requires that the designer knows how agents' expectations of a moment of the opponents' types moves, conditional on their own type, and that this is described by a function that is nowhere constant. This result, which arises discontinuously as generalized independence is lifted, is somewhat reminiscent of the full surplus extraction (FSE) results (Crémer and McLean, 1985, 1988), which also arise discontinuously in Bayesian environments, when minimal degrees of correlation are introduced. Importantly, however, FSE does *not* generally ensue in our setup. If the belief-restrictions are not Bayesian, even if any d can be implemented under the responsive moment condition, there may still be bounds to the surplus that can be extracted (Propositions 2 and 3). Information rents generally remain, and their size depends on the joint properties of the allocation rule, agents' preferences, and the belief restrictions. Moreover, information rents shrink as the belief sets get finer, and the designer relies on more information about agents' beliefs (Prop. 5). At the extreme, if $\mathcal{B}$ is a Bayesian setting with correlated types, then FSE obtains. In fact, under a novel 'full rank' condition, we provide the following 'any*thing* goes' result (Proposition 4): in a Bayesian setting that satisfies 'full rank', for any (d, t), there exist transfers t' that are both incentive compatible and that attain the same expected payments as t.

Jointly, these results highlight an important feature of our framework. Specifically, since their very inception, FSE results have famously been received as disturbing (cf. footnote 1). In response, mechanism design has largely shied away from studying environments with correlated or, in the terminology of Postlewaite and Schmeidler (1986), 'non-exclusive information'. But the economic relevance of these settings can hardly be underplayed, and their analysis should not be put aside, merely due to the inadequacy of the classical theoretical toolbox.[6] Our results show that the *belief-restrictions* framework is capable of expressing

[5] Note that both belief-free settings (where $B_{\theta_i} = \Delta(\Theta_{-i})$ for all i and θ_i) and Bayesian settings with independent types (where, for each i, all B_{θ_i} consist of the same singleton for all θ_i) are special cases of *generalized independence*. In the latter case, Myerson's (1981) revenue equivalence Theorem obtains from Corollary 1 as a special case. In Section 5 we show that in fact, a *robust* version of *revenue equivalence* holds whenever the belief restrictions are such that $\cap_{\theta_i \in \Theta_i} B_{\theta_i} \neq \emptyset$ for all i and θ_i (Corollary 6).

[6] Again, in the words of Crémer and McLean (1988): "[...] we should stress that in our opinion the independence assumption should be used only with great caution [...]. It does enable the derivation of

a meaningful notion of non-exclusive information, where we show that even a minimalist notion of *comovement* between types and beliefs may serve as an additional tool to screen types and attain implementation, while at the same time avoiding the pitfalls of FSE results, which instead really hinge upon full and precise knowledge of the distribution that describes agents' beliefs. This framework may thus help mechanism design's reappropriation of environments with non-exclusive information, in which distilling intuitive and reliable economic intuition has long appeared elusive, within the prevailing paradigm.

In Section 5 we discuss further implications of our main Theorems, and their connections with known results in the literature. We also expand on other uses of belief-based terms, including in environments with *generalized independence* (such as Bayesian settings with independent types), where they cannot be used to improve on the canonical transfers' ability to attain incentive compatibility, but where they may still be useful to pursue extra desiderata, beyond incentive compatibility.[7] Propositions 6 and 7, for instance, provide both possibility and impossibility results for unique implementation via moment conditions, to gain perspective on what kind of information is useful for the design of transfers for unique implementation, as a function of the agents' preferences and allocation rule. We also discuss further methodological considerations. Theorem 4, in particular, provides a characterization of the equilibrium payoffs that clarifies the connection between standard envelope formulae and the belief-based terms at the center of our analysis, and to compare the relative merits of the envelope approach and of the *first-order approach* that we pursued in this paper. Section 6 discusses the related literature. Section 7 concludes.

2 Framework

Payoff Environments. The payoff environment represents agents' information about everyone's preferences over the set of feasible allocations, and an allocation rule that maps agents' information to the space of allocations, and which represents the designer's objective. Formally, let $I = \{1, ..., n\}$ denote the (finite) set of agents, $X \subseteq \mathbb{R}^m$ the set of allocations. For each $i \in I$, we let Θ_i denote the set of player i's payoff types, with typical element θ_i, assumed private information. We adopt the standard notation for type profiles, and let $\theta \in \Theta := \times_{i \in I} \Theta_i$, and for each i, we let $\theta_{-i} \in \Theta_{-i} := \times_{j \neq i} \Theta_j$. For each i, the *valuation function* is denoted $v_i : X \times \Theta \to \mathbb{R}$. Note that we allow v_i to depend on the entire profile of types, so as to allow the case of interdependent values. For each i, we let $t_i \in \mathbb{R}$ denote

results that on the surface look more 'realistic' (there is no full extraction of the surplus). However, the derivation of these results rely on a very 'unrealistic' assumption. Furthermore, [...] a small deviation from this assumption can induce fundamentally different results.".

[7] Several such 'extra desiderata' have been considered in the literature (such as budget balance (d'Aspremont and Gérard-Varet, 1979), surplus extraction (Crémer and McLean, 1985, 1988), collusion-proofness (Laffont and Martimort, 1997; Che and Kim, 2006; Safronov, 2018), stability (Mathevet, 2010; Mathevet and Taneva, 2013; Healy and Mathevet, 2012; Sandholm, 2002, 2005, 2007), uniqueness (Ollár and Penta, 2017, 2022, 2023), etc.). But, prior to this paper, the lack of a general characterization of the incentive compatible transfers, these analyses have escaped a unified, systematic analysis.

the monetary transfer to agent i, and assume that i's utility for each $(x,t) \in X \times \mathbb{R}^n$, given type profile $\theta \in \Theta$, is equal to $u_i(x,t,,\theta) = v_i(x,\theta) + t_i$. The model can thus accommodate both private and interdependent values, as well as general externalities in consumption, including the cases of pure private goods and public goods. An *allocation rule* is a function $d : \Theta \to X$, which assigns, to each type profile, the allocation that the designer wishes to implement. We maintain throughout the following assumptions:

Assumption 1 (Payoff Environment)**.** $((\Theta_i, v_i)_{i\in I}, d)$ *is such that for all* $i \in I$*:*

(i) $\Theta_i := [\underline{\theta}_i, \overline{\theta}_i] \subset \mathbb{R}$

(ii) v_i *is twice continuously differentiable*

(iii) d *is piecewise differentiable.*[8]

Note that these assumptions require that d is only *piecewise* differentiable in types, and hence the model also accommodates discontinuous allocation rules, which are common for instance in auctions, bilateral trade and assignement problems. The main substantial restriction is the one-dimensionality of the payoff types.[9]

Belief Restrictions. We model the maintained assumptions on agents' beliefs via the belief-restrictions we first introduced in Ollár and Penta (2017). We let $\Delta(\Theta_{-i})$ denote the set of probability measures over Θ_{-i}, which represent beliefs about the opponents' types. A belief restriction is a collection of sets of possible beliefs, for each type of each agent, over the set of type profiles of the other agents. Formally, a *belief restriction* is a collection $\mathcal{B} = ((B_{\theta_i})_{\theta_i\in\Theta_i})_{i\in I}$, such that, $B_{\theta_i} \subseteq \Delta(\Theta_{-i})$ is non-empty for each i and θ_i. Belief restrictions can be used to accommodate varying degrees of robustness. For instance:

(i) The *belief-free settings* of the early literature on robust mechanism design (e.g., Bergemann and Morris (2005, 2009a,b), Penta (2015), Müller (2016), etc.) are obtained by letting $B_{\theta_i} = \Delta(\Theta_{-i})$ for all i and $\theta_i \in \Theta_i$, and denoted by $\mathcal{B}^{BF} = ((B^{BF}_{\theta_i})_{\theta_i\in\Theta_i})_{i\in I}$.

(ii) The standard *Bayesian settings* are obtained if the belief restrictions are commonly known and each belief set is a singleton for every type: $B^{\diamond}_{\theta_i} = \{b^{\diamond}_{\theta_i}\}$ for all i and $\theta_i \in \Theta_i$. In this case, each player's payoff type uniquely pins down the infinite belief hierarchy, as in the interim formulation in a standard Harsányi type space. Further, in the special case of a *common prior* type space, there exists $p \in \Delta(\Theta)$ s.t., for each i and $\theta_i, p(\cdot|\theta_i) = b^{\diamond}_{\theta_i} \in \Delta(\Theta_{-i})$. If, furthermore, such a common prior is *independent* across agents, then we also have $b^{\diamond}_{\theta_i} = b^{\diamond}_{\theta'_i}$ for all $\theta_i, \theta'_i \in \Theta_i$ and for all $i \in I$.

(iii) Intermediate notions of robustness obtain whenever $B_{\theta_i} \subset \Delta(\Theta_{-i})$ for some θ_i, but not all singletons. Special cases have been considered, for instance, to model that

[8] We say that $f : S \to \mathbb{R}$ is *piecewise differentiable* on a closed and convex set $S \subset \mathbb{R}^n$ if there exist a collection $(S_k)_{k=1,\dots,K}$ of pairwise disjoint convex sets such that $\cup_{k=1}^K S_k = S$, and continuously differentiable functions $g_k : S \to \mathbb{R}$, $k = 1 \dots K$, such that $f = \sum_{k=1}^K f_k$ where, for each $k = 1, ..., K$, $f_k(x) = \mathbf{1}_{[x\in S_k]} \cdot g_k(x)$. We say that $f : S \to \mathbb{R}^m$ is piecewise differentiable if it is componentwise piecewise differentiable.

[9] It is well known that incentive compatibility is significantly more problematic outside of this domain, as multidimensionality of types severaly limits its possibility (Jehiel and Moldovanu, 2001 and Jehiel et al., 2006). This approach is extended to the multidimensional case in Ollár (2024).

agents commonly know some moments of the distributions of the opponents' types (*common knowledge of moment conditions*, Ollár and Penta, 2017), or that the opponents' types are identically distributed (*common belief in identicality*, Ollár and Penta, 2023).

These are examples of special instances from the mechanism design literature, but the framework is more general. We stress that since the focus here is on partial implementation and incentive compatibility, the results in this paper do not require the belief restrictions to be common knowledge among the agents: they depend only on the *first-order beliefs*.

Given belief restrictions $\mathcal{B} = ((B_{\theta_i})_{\theta_i \in \Theta_i})_{i \in I}$ and $\mathcal{B}' = ((B'_{\theta_i})_{\theta_i \in \Theta_i})_{i \in I}$, we write $\mathcal{B} \subseteq \mathcal{B}'$ to denote that $B_{\theta_i} \subseteq B'_{\theta_i}$ for all $i \in I$ and all $\theta_i \in \Theta_i$. If $\mathcal{B} \subseteq \mathcal{B}'$, then $\mathcal{B}$ imposes stronger restrictions than $\mathcal{B}'$, in that the designer can rule out more beliefs in the former than in the latter. In this sense, the belief-free model $\mathcal{B}^{BF}$ is minimal in the information that the designer has, as any model $\mathcal{B}$ is such that $\mathcal{B} \subseteq \mathcal{B}^{BF}$. At the opposite extreme, any Bayesian setting $\mathcal{B}^{\diamond}$ is maximal, as no distinct belief restriction $\mathcal{B}$ is such that $\mathcal{B} \subseteq \mathcal{B}^{\diamond}$. Belief restrictions $\mathcal{B}^{id}$ are an example of an intermediate robustness requirement, $\mathcal{B}^{\diamond} \subseteq \mathcal{B}^{id} \subseteq \mathcal{B}^{BF}$.

Mechanisms. A mechanism is a tuple $\mathcal{M} = ((M_i)_i, g)$, where M_i denotes the set of messages of player i, and $g : M \to X \times \mathbb{R}^n$ assigns to each profile of messages, $m \in M := \times_{i \in I} M_i$, an allocation and a profile of payments. In a direct mechanism agents report their type (i.e., $M_i = \Theta_i$ for all i) and the allocation is chosen according to d (i.e. $g(m) = (d(m), t(m))$). A *direct mechanism* therefore is completely pinned down by the *transfer scheme* $t = (t_i)_{i \in I}$, where for each $i \in I$, $t_i : M \to \mathbb{R}$ specifies the transfer to agent i for all profile of reports $m \in M \equiv \Theta$. Each (direct) mechanism (d, t) induces an incomplete information game with ex-post payoff functions $U_i^t(m; \theta) = v_i(d(m), \theta) + t_i(m)$, which under the maintained assumptions are bounded. We adopt the following notation: For any $\theta_i \in \Theta_i$, $b \in \Delta(\Theta_{-i})$ and $m_i \in M_i$, we let $\mathbb{E}^b U_i^t(m_i; \theta_i) := \int_{\Theta_{-i}} U_i^t(m_i, \theta_{-i}; \theta_i, \theta_{-i})\, db$, and for any $f : \Theta \to \mathbb{R}$, $\theta_i \in \Theta_i$ and $b \in B_{\theta_i}$, we let $\mathbb{E}^b[f(\theta_i, \theta_{-i})] := \int_{\Theta_{-i}} f(\theta_i, \theta_{-i})\, db$.

2.1 Incentive Compatibility

Incentive compatibility requires that truthtelling be a mutual best response, for all beliefs that are consistent with the belief restrictions $\mathcal{B}$.

Definition 1. *A direct mechanism (d, t) is **$\mathcal{B}$-incentive compatible** ($\mathcal{B}$-IC) if for all $i \in I$, $\theta_i \in \Theta_i$, $m_i \in M_i$, $\mathbb{E}^b U_i^t(m_i; \theta_i) \leq \mathbb{E}^b U_i^t(\theta_i; \theta_i)$ for all $b \in \mathcal{B}_{\theta_i}$.*

When d is clear from the context, we say that the transfer scheme t is $\mathcal{B}$-IC.

Here we take this notion as given, but it is straightforward to show that if robustness concerns are embedded within the equilibrium notion in a natural way, via a suitable generalization of ex-post and Bayes-Nash equilibrium, then a standard *revelation principle* obtains, given which the restriction to direct mechanisms and incentive compatibility is

without loss.[10] With respect to this approach to robustness, which is in complete analogy with ex-post equilibrium in belief-free settings, as well as with Bewley (2002)'s model of incomplete preferences under uncertainty, the notion of $\mathcal{B}$-IC is without loss.

Alternatively, one could approach robustness by requiring implementation on all type spaces consistent with the belief restrictions. From this perspective, it can be shown that an allocation rule $d : \Theta \to X$ is interim (or Bayesian) incentive compatible on *all* type spaces consistent with the belief restrictions if and only if d is $\mathcal{B}$-IC in the sense above (cf. Ennuschat and Penta, 2025). Hence, as far as implementation of such allocation rules goes, nothing would be gained by considering mechanisms that also elicit agents' beliefs.[11]

In Bayesian environments, $\mathcal{B}$-IC is equivalent to interim (or Bayesian) incentive compatibility (IIC). At the opposite extreme, in belief-free settings, it is equivalet to ex-post incentive compatibility (ep-IC). For intermediate belief restrictions, $\mathcal{B}$-IC is weaker than ep-IC (since truthful revelation need not be optimal for all beliefs about Θ_{-i}) but it is stronger than IIC (in that it requires truthful revelation to be optimal for all beliefs in B_{θ_i}, not just for one). More generally: If $\mathcal{B} \subseteq \mathcal{B}'$, and (d, t) is $\mathcal{B}'$-IC, then it is also $\mathcal{B}$-IC.

2.2 The canonical transfers

For later reference, we recall the definition of the *canonical transfers*, $t^* = (t_i^*(\cdot))_{i \in I}$, which are pinned-down by the first-order conditions for ep-IC. Formally, for each i and m,

$$t_i^*(m) = -v_i(d(m), m) + \int_{\underline{\theta}_i}^{m_i} \frac{\partial v_i}{\partial \theta_i}(d(s_i, m_{-i}), s_i, m_{-i})\, ds_i. \tag{1}$$

We will refer to (d, t^*) as the canonical *direct* mechanism, and let $U_i^* : M \times \Theta \to \mathbb{R}$ denote the ex-post payoff function associated to it. For later reference, we provide some well-known results about the canonical transfers:[12]

Lemma 1. *(i) If (d, t) is ep-IC, then for each i there exists $\kappa_i : M_{-i} \to \mathbb{R}$ such that for each $t_i(m) = t_i^*(m) + \kappa_i(m_{-i})$.*

[10] Formally, given an arbitrary mechanism $\mathcal{M}$ and belief restrictions $\mathcal{B}$, a strategy profile $(\sigma_i)_{i \in I}$, where $\sigma_i : \Theta_i \to M_i$ for each i, is a *$\mathcal{B}$-equilibrium* if, for all i, θ_i and $b \in B_{\theta_i}$, $\sigma_i(\theta_i) \in \operatorname{argmax}_{m_i \in M_i} \int_{\Theta_{-i}} u_i(m_i, \sigma_{-i}(\theta_{-i}); \theta_i, \theta_{-i})\, db$. Ex-post and Bayes-Nash equilibrium obtain, respectively, for the special cases of $\mathcal{B}^{BF}$ and Bayesian belief restrictions. It is straightforward to show that a revelation principle also holds in this case. That is: d is $\mathcal{B}$-IC if and only if there exists a (not necessarily direct) mechanism $\mathcal{M} = (M_i, g_i)_{i \in I}$ and a $\mathcal{B}$-equilibrium σ of $\mathcal{M}$ such that $g(\sigma(\theta)) = d(\theta)$ for all $\theta \in \Theta$.

[11] This is not to say that *all* interesting mechanism design questions necessarily boil down to implementing some $d : \Theta \to X$, and hence satisfy the 'separability' condition in Ennuschat and Penta (2025). Without separability, the restriction to $\mathcal{B}$-IC may be with loss, but those cases are beyond the scope of this paper.

[12] The 'canonical transfers', and the associated *canonical direct mechanism* (d, t^*), should not be confused with the 'canonical mechanism', which traditionally refers to Maskin's (non-direct) mechanism for *full* implementation. Special instances of the canonical direct mechanism have appeared throughout the literature on *partial* implementation, e.g. in the auction mechanisms of Myerson (1981), Dasgupta and Maskin (2000) and Segal (2003), the pivot mechanisms of Milgrom (2004) and Jehiel and Lamy (2018), the public goods mechanisms in Green and Laffont (1977), Laffont and Maskin (1980) and Jehiel and Moldovanu (2001).

(ii) If valuations satisfy the ex-post Single-Crossing Condition *(ep-SCC: $\partial^2 v_i/\partial x \partial \theta_i > 0$ for all i), then (d, t^*) is ep-IC if and only if d is increasing in each θ_i.*

(iii) If d is increasing, then (d, t^*) is ep-IC if and only if $\partial^2 v_i/\partial x \partial \theta_i \geq 0$ for all i.

Point (i) shows that the canonical transfers are essentially the only ones that achieve ep-IC, if at all possible. The results in points (ii) and (iii) follow from the fact that, with t^* pinned-down by the first-order conditions for ep-IC, single-crossing and monotonicity ensure that U_i^* also satisfies the second-order conditions.

2.3 Leading Example and Preview of Results

Example 1 (Implementing a Policy under Opposing Interests)**.** A government is deciding on the quantity $x \geq 0$ of spending in pollution reduction activities. Society consists of two agents, and the government's desired level of expenditure is $d(\theta) = K(\theta_1 + \theta_2)$, where $K > 0$, and $\theta_i \in [0, 1]$ denotes the productivity of agent i, which is their private information. Agents work in different sectors, with opposing preferences over pollution reduction, as a function of their productivity: their valuation functions are $v_1(\theta, x) = \theta_1 x$ and $v_2(\theta, x) = -\theta_2 x$, respectively.[13] Due to the agents' opposing interests, the single-crossing and monotonicity conditions from Lemma 1 fail in this setting, and hence implementation would be impossible in a belief-free sense. To see this, note that the formula for the canonical transfers (eq. (1)) in this example induces payoff functions $U_1^*(m, \theta) = \theta_1 K(m_1 + m_2) - K\frac{1}{2}m_1^2$ and $U_2^*(m, \theta) = -\theta_2 K(m_1 + m_2) + K\frac{1}{2}m_2^2$. Hence, while truthful revelation satisfies the F.O.C. for both agents, since the allocation rule moves with θ_2 in the opposite direction of 2's marginal utility for x, U_2^* is convex in m_2 and hence the S.O.C. fail for agent 2. The canonical transfers therefore are not ep-IC.

Now suppose that the designer knows that both agents' expect the opponent's type, on average, to be half of their own. But the actual distributions that describe their beliefs are not known to the designer. Formally, $\mathcal{B}$ is such that $B_{\theta_i} = \{b \in \Delta(\Theta_j) : \mathbb{E}^b(\theta_j) = \theta_i/2\}$ for each i and θ_i. Under the canonical transfers, truthful revelation violates the necessary S.O.C. also with respect to the interim payoffs, for all beliefs consistent with $\mathcal{B}$, and hence (d, t^*) is not $\mathcal{B}$-IC either. But now consider these modified transfers for agent 2:

$$t_2^{mod}(m) = t_2^*(m) + \underbrace{-K\left(m_2^2 - 4m_1 m_2\right)}_{\text{"belief-based term": } \beta_2(m)}, \tag{2}$$

which induce ex-post payoffs $U_2^{mod}(m; \theta) = U_2^*(m; \theta) + \beta_i(m)$. Note that the belief-based term is such that, for each θ_2, and for each $b \in B_{\theta_2}$:

[13] Clearly, this policy is not efficient in this example. This may be due to political reasons that may lead the government to favor a particular agenda, despite the opposite preferences of certain social groups, or because the policy reflects environmental externalities that are not accounted for by agents' preferences.

$$\mathbb{E}^b\left[\frac{\partial \beta_2}{\partial m_2}(\theta_1,\theta_2)\right] = -K \cdot \underbrace{\mathbb{E}^b[(2\theta_2 - 4\theta_1)]}_{=0 \text{ under } \mathcal{B}} = 0. \tag{3}$$

Furthermore, we also have $\mathbb{E}^{b_{\theta_2}}\left[\frac{\partial^2 \beta_2}{\partial^2 m_2}(\theta)\right] = -2K$, which implies $\mathbb{E}^{b_{\theta_i}}\left[\frac{\partial^2 U_i^{mod}}{\partial^2 m_i}(\theta)\right] = -K < 0$. Hence, truthful revelation satisfies both the first- and second-order conditions, and hence (d, t^{mod}) is $\mathcal{B}$-IC. □

Note that the transfers in (2) can be written as $t_i^{mod}(m) = t_i^*(m) + \beta_i(m)$, where $\beta_i : M \to \mathbb{R}$ is a *belief-based term* that satisfies $\mathbb{E}^{b_{\theta_i}}\left[\frac{\partial \beta_i}{\partial m_i}(\theta_i, \theta_{-i})\right] = 0$ for all θ_i and $b_{\theta_i} \in B_{\theta_i}$. Theorem 1 in Section 3.1 shows that this holds in general: in any envrionment, and for any belief-restrictions $\mathcal{B}$, any $\mathcal{B}$-IC transfers must take this form.[14] Theorem 2 in Section 3.2 shows that, in order to guarantee that the second-order conditions are satisfied, the belief-based terms should also be such that $\mathbb{E}^b\left[\frac{\partial^2 U_i^*}{\partial^2 m_i}(\theta_i, \theta_{-i})\right] < -\mathbb{E}^b\left[\frac{\partial^2 \beta_i}{\partial^2 m_i}(\theta_i, \theta_{-i})\right]$ for all θ_i and $b \in B_{\theta_i} \subseteq \Delta(\Theta_{-i})$. Theorem 3 in Section 3.3 provides a characterization that highlights the role of belief-based terms in overcoming failures of standard single-crossing and monotonicity conditions.

Overall, these results highlight a general design principle, that reduces the design of incentive compatible transfers to the design of suitably restricted *belief-based terms*. These results serve several purposes. They guide the design of transfers to overcome violations of standard single-crossing and monotonicity conditions, as we illustrated in Example 1. They also enable a characterization of the set of $\mathcal{B}$-IC transfers, which is useful to pursue other questions or objectives beyond incentive compatibility. For instance, in the setting of Example 1, in Section 3.2 we characterize the set of $\mathcal{B}$-IC transfer schemes, we identify those that minimize the cost of implementation for the designer, as well as those that attain *unique* implementation. We also show that, in this setting, incentive compatibility always grants informational rents to the agents. Hence, although these belief restrictions significantly expand the possibility of implementation, in a setting where standard single-crossing and monotonicity conditions do not hold, information rents still remain, and there are bounds on the extent to which the designer can extract surplus from the agents.[15]

These insights are generalized in several directions by the results in Section 4. In particular, we show that under a weak property of 'comovement' between types and beliefs, *any* allocation rule can be implemented (Proposition 1). Yet, unless the environment is Bayesian, informational rents generally remain (Proposition 2-4), and they get larger as the robustness requirements get stronger (Propoposition 5).

[14] In Section 5.1 we discuss several implications of this result, including a *robust* version of the *revenue equivalence theorem*, which we obtain under a notion of *generalized independence* for non-Bayesian settings.

[15] This would not be the case in a Bayesian setting. For instance, if the designer knows that each type θ_i's beliefs about the opponent is that their types are distributed according to a uniform over $[0, \theta_i]$ (which is consistent with the belief restrictions above), then not only could d be implemented, but the transfers could be adjusted so as to extract the full surplus from the agents (Proposition 4).

3 Main Results

In this section we provide the main results of our paper. Theorem 1 in Section 3.1 focuses on the necessary implications of $\mathcal{B}$-incentive compatibility; Theorem 2 in Section 3.2 also provides sufficient conditions in environments with differentiability. Theorem 3 provides a full characterization in general environments. Throughout this Section, we illustrate the applicability of these results within our running Example. In Section 4 we apply them to obtain permissive implementation results in general environments (with or without single-crossing and monotonicity), and we formalize the sense in which our framework enables us to capture a meaningful notion of 'comovement' of beliefs and types that is useful for implementation, but without incurring into the pitfalls of 'full-surplus extraction' results.

3.1 $\mathcal{B}$-IC Transfers: Necessity

In this section we derive necessary conditions for $\mathcal{B}$-IC transfers, for general belief restrictions. Our results are based on a generalization of the classical *first-order approach*, that identifies necessary conditions for *local* incentive compatibility constraints (cf. Rogerson, 1985; Jewitt, 1988). Compared to the classical results, the main difference is that, instead of focusing on the ex-post payoff function, we take an interim perspective and consider the expected payoff function of every type θ_i, for all beliefs in the set B_{θ_i}.

Theorem 1 ($\mathcal{B}$-IC Transfers (Necessity))**.** *Under the maintained assumptions, if t is piecewise differentiable and (d,t) is $\mathcal{B}$-IC, then for all i, and for all $m \in M \equiv \Theta$,*

$$t_i(m) = t_i^*(m) + \beta_i(m), \tag{4}$$

where $\beta_i : M \to \mathbb{R}$ is piecewise differentiable; and it is such that for all θ_i and for all beliefs $b \in B_{\theta_i}$ with a piecewise differentiable pdf, at all points of differentiability,

$$\left.\frac{\partial \mathbb{E}^b\left[\beta_i(m_i, \theta_{-i})\right]}{\partial m_i}\right|_{m_i=\theta_i} = 0. \tag{5}$$

Hence, in order to design a $\mathcal{B}$-IC transfer scheme, it is without loss to restrict attention to additive modifications of the canonical transfers, provided that the added terms satisfy the expectation condition in Equation (5). We refer to the functions $\beta_i : M \to \mathbb{R}$ that satisfy Equation (5) as the *belief-based terms that are consistent with $\mathcal{B}$* (or simply *belief-based terms*, when $\mathcal{B}$ is clear from the context). In hindsight, the result may perhaps appear straightforward for the special case where everything is differentiable (see the introduction). But it remains true under the general maintained assumptions of the previous section. Notwithstanding the relative simplicity of the result, Theorem 1 has a rich set of implications, for both negative and positive results. First, consider the following definition:

Definition 2. *$\mathcal{B}$ satisfies **generalized independence** if the belief sets are constant across types: for all $i \in I$, $B_{\theta_i} = B_{\theta_i'}$ for all $\theta_i, \theta_i' \in \Theta_i$.*

Note that this condition holds in any of the following special cases: *belief-free* settings; Bayesian models with *independent types*; the $\mathcal{B}^{id}$-restrictions for *common belief in identicality* from Ollár and Penta (2023). With this, Theorem 1 implies the following:

Corollary 1. *Fix* (v, d). *Assume that* $\mathcal{B}$ *satisfies* generalized indepence. *If there exist a* $\mathcal{B}$*-IC transfer scheme* t*, then* t^* *is* $\mathcal{B}$*-IC.*

In words, under *generalized independence*, an allocation rule is implementable if and only if the direct canonical mechanism is $\mathcal{B}$-IC.[16] Hence, unlike in our leading example, the information about beliefs in these settings cannot be used to establish incentive compatibility when the canonical transfers fail it. Intuitively, if all types of an agent share the same belief sets, beliefs are not helpful to screen types in a robust way, beyond what can be achieved based on the ex-post payoffs. Further implications of Theorem 1 will be discussed in Section 5, where we also present a *robust revenue equivalence* result that holds under a weaker notion of generalized independence.

3.2 Incentive Compatible Transfers in the Differentiable Case

By design, the transfers that satisfy the conditions in Theorem 1 are such that truthful-revelation satisfies the *first-order conditions* of the interim payoff functions, for all beliefs consistent with the belief restrictions. Fully understanding incentive compatibility however also requires ensuring that the payoff functions have the right curvature. This is typically what single-crossing and monotonicity conditions do. In the absence of these conditions, the next result shows how the belief-based terms can be used to induce concavity of the payoff functions in settings with differentiability. This assumption will be relaxed in Theorem 3, which provides a general characterization that highlights the role that belief-based terms play in overcoming failures of standard single-crossing and monotonicity conditions.

Theorem 2 (Conditions under Differentiability)**.** *Assume that* v_i, t_i, d *are all twice differentiable, and for each* i*, let* $\beta_i := t_i - t_i^*$.

Necessity: *Transfers* t *are* $\mathcal{B}$*-IC only if for all* i, θ_i *and for all* $b \in B_{\theta_i}$

(i) $\mathbb{E}^b[\partial_i \beta_i(\theta_i, \theta_{-i})] = 0$ *and*

(ii) there exists an open neighborhood of θ_i*,* $\mathcal{N}_{\theta_i}$ *such that for all* $m_i \in \mathcal{N}_{\theta_i}$*:*

$$\mathbb{E}^b[\partial_{ii}^2 U_i^*(m_i, \theta_{-i}; \theta_i, \theta_{-i})] \leq -\mathbb{E}^b[\partial_{ii}^2 \beta_i(m_i, \theta_{-i})]. \tag{6}$$

Sufficiency:*: Transfers* t *are* $\mathcal{B}$*-IC if for all* i, θ_i *and for all* $b \in B_{\theta_i}$*, Condition (i) holds and Inequality* (6) *holds for all* $m_i \in M_i$.

[16] This result generalizes the first point of Lemma 1, which obtains from this Corollary for the special case of belief-free restrictions. It also generalizes Theorem 1 in Ollár and Penta (2023), which only referred to the special case of $\mathcal{B}^{id}$-restrictions. These and other related results will be further discussed in Section 5.1.

Condition (i) states the necessary condition from Theorem 1, for the differentiable case; Condition (ii) states the necessary second-order condition instead, which relates the curvature of the payoff function of the canonical direct mechanism to the belief-based term. If they hold globally, then these conditions are also sufficient.

In its simplicity, Theorem 2 distills a general design principle. To see this, note that the canonical transfers are $\mathcal{B}$-IC if the term on the left-hand side of (6) is less than zero. When this is not the case, the belief-based term can be used to relax this constraint: if belief-based terms exist that satisfy Condition (i), and that are sufficiently concave so as to make (6) hold for all m_i, then $\mathcal{B}$-IC can be attained. The idea therefore is to identify sufficiently concave belief-based terms, subject to Condition (i) being satisfied. This is useful both to recover incentive compatibility when the canonical transfers do not achieve it (like we did, for instance, in Ex. 1), but also to identify the limits of $\mathcal{B}$-IC, as we illustrate next:

Example 2 (Ex. 1, continued)**.** To characterize the set of $\mathcal{B}$-IC transfers, under belief restrictions $\mathcal{B}$ s.t. $B_{\theta_i} = \{b \in \Delta(\Theta_j) : \mathbb{E}^b(\theta_j) = \theta_i/2\}$ for all θ_i and i, first we identify the set of belief-based terms that satisfy the necessary condition in part 1 of Theorem 2. (We maintain in this example that the lowest type of each agent always pays 0.) In this setting, $\beta_i : M \to \mathbb{R}$ satisfies such condition if and only if $\partial_i \beta_i (m_i, m_j) = (m_i - 2m_j) H_i (m_i)$ where H_i is a real function on $M_i \equiv \Theta_i$ (see the Appendix). Hence, belief-based terms in this setting must necessarily take the following form:

$$\beta_i(m) = \int_0^{m_i} (s - 2m_j) H_i(s) ds$$

Notice that, since for each θ_i and $b \in B_{\theta_i}$ we have $\mathbb{E}^b[\theta_j] = \theta_i/2$ the following simplification occurs for all such beliefs:

$$\partial^2_{ii} \mathbb{E}^b[\beta_i (\theta_1, \theta_2)] = H_i(\theta_i) + \left(\theta_i - 2\mathbb{E}^b[\theta_j|\theta_i]\right) H_i'(\theta_i) = H_i(\theta_i)$$

Given this, for agent 1 part 2 of Theorem 2 holds if and only if, for all beliefs consistent with the belief-restrictions, $-K + \partial^2_{11} \mathbb{E}^b[\beta_1 (\theta_1, \theta_2)] \leq 0$. Exploiting the condition above, this simplifies to $H_1(\theta_1) \leq K$ for all θ_1. Similarly, for agent 2 we obtain $H_2 (\theta_2) \leq -K$ for all θ_2. Hence, a transfer scheme is $\mathcal{B}$-IC if and only if it takes the form

$$\begin{aligned} t_1 (m_1, m_2) &= -\frac{K}{2} m_1^2 + \int_0^{m_1} (s - 2m_2) H_1 (s) \, ds, \text{ and} \\ t_2 (m_1, m_2) &= \frac{K}{2} m_2^2 + \int_0^{m_2} (s - 2m_1) H_2 (s) \, ds, \end{aligned}$$

subject to the restriction on the H_i functions above. Exploiting again the fact that, for

each θ_i and $b \in B_{\theta_i}$, $\mathbb{E}^b[\theta_j] = \theta_i/2$, the expected transfers at the truth-telling profile are:

$$\begin{aligned}\mathbb{E}^b[t_1(\theta)|\theta_1] &= -\frac{K}{2}\theta_1^2 + \int_0^{\theta_1}(s-\theta_1)H_1(s)\,ds, \text{ and}\\ \mathbb{E}^b[t_2(\theta)|\theta_2] &= \frac{K}{2}\theta_2^2 + \int_0^{\theta_2}(s-\theta_2)H_2(s)\,ds,\end{aligned}$$

which are minimized by setting each $H_i(\theta_i)$ at the corresponding upper bound, that is $H_1 \equiv K$ and $H_2 \equiv -K$. The resulting transfers, $t_1^{Cmin}(m_1,m_2) = -2Km_2m_1$, and $t_2^{Cmin}(m_1,m_2) = 2Km_1m_2$, therefore attain the lowest expected transfers to each agent pointwise, for each type realization $\theta \in \Theta$ and regardless of agents' true beliefs within B_{θ_i}.

If, instead of cost-minimization, the designer wished to achieve *unique* implementation (which is not attained by t^{Cmin}), then the H-terms should be chosen in order to ensure weak *strategic externalities* (Ollár and Penta, 2023, 2025). Within the $\mathcal{B}$-IC constraints, these are minimized by setting $H_1 \equiv 0$ and $H_2 \equiv -2K$. The resulting transfers, $t_1^{unique}(m_1,m_2) = -\frac{K}{2}m_1^2$, and $t_2^{unique}(m_1,m_2) = -\frac{K}{2}m_2^2 + 4Km_1m_2$, induce truth-telling as the unique strategy that survives two rounds of dominance under the belief restrictions. But setting $H_1 \equiv 0$ $H_2 \equiv -K-\varepsilon$ for some arbitrarily small $\varepsilon > 0$ ensures both uniqueness and implementation costs that are arbitrarily close to the minimal ones. □

The insights in this example will be generalized in several directions by the results in Section 4. In particular, we will show that under a weak property of 'comovement' between types and beliefs, then *any* allocation rule can be implemented. Yet, unless the environment is Bayesian, information rents in general remain, just as in the example above.

3.3 The general case: A Full Characterization

We provide next a characterization of the $\mathcal{B}$-IC transfers in general environments, that highlights the role that belief-based terms may play in overcoming failures of standard single-crossing and monotonicity conditions, as it was the case in the previous example.

Theorem 3 ($\mathcal{B}$-IC: Characterization)**.** *Under the maintained assumptions of Theorem 1, for each i, let $\beta_i := t_i^* - t_i$. Then, (d,t) is $\mathcal{B}$-IC if and only if for all i, θ_i, $b \in B_{\theta_i}$ and m_i:*

$$\mathbb{E}^b\left[\int_{m_i}^{\theta_i}\left(\tfrac{\partial v_i}{\partial \theta_i}(d(s,\theta_{-i}),s,\theta_{-i}) - \tfrac{\partial v_i}{\partial \theta_i}(d(m_i,\theta_{-i}),s,\theta_{-i})\right)ds\right] \geq \mathbb{E}^b\Big[\beta_i(m_i,\theta_{-i}) - \beta_i(\theta)\Big].$$

To understand this result, let us first consider the *belief-free* case, where $\mathcal{B}$-IC coincides with ep-IC. First, as this condition must hold for all beliefs, it must also hold in the ex-post sense, and hence we can just focus on the terms inside the square brackets. Second, as discussed, in belief-free settings the necessary condition in Theorem 1 implies that the belief-based terms are constant in own message, and hence the right-hand side of the conditions in Theorem 3 are equal to zero. Thus, for belief-free settings, the following holds:

Corollary 2 (ep-IC and ep-SCM)**.** *Under the maintained assumptions of Theorem 1, , (d, t^*) is ep-IC if and only if for all θ_i, θ_i' and for all θ_{-i}:*[17]

$$\left[\frac{\partial v_i}{\partial \theta_i}\left(d\left(\theta_i', \theta_{-i}\right), \theta_i, \theta_{-i}\right) - \frac{\partial v_i}{\partial \theta_i}\left(d\left(\theta_i, \theta_{-i}\right), \theta_i, \theta_{-i}\right)\right] \cdot \left(\theta_i' - \theta_i\right) \geq 0.$$

This condition entails joint restrictions on the single-crossing properties of the valuation functions, and on the monotonicity of the allocation rule. The known results in Lemma 1 (points (ii) and (iii)), in particular, follow immediately from this Corollary. For these reasons, we refer to this condition as *ex-post Single-Crossing and Monotonicity* (ep-SCM).

Analogously, in a Bayesian setting with independent types, the same logic implies that IIC is possible if and only if a suitable *interim*-SCM condition is satisfied:

Corollary 3 (IIC with Independent Types)**.** *Let $\mathcal{B}^\diamond$ be a Bayesian environment with independent types, and let $b_i^\diamond \in \Delta(\Theta_{-i})$ denote agent i's beliefs, regardless of his type. Then, under the maintained assumptions of Theorem 1, an IIC transfer scheme exists if and only if for all i, and for almost all pairs of θ_i, θ_i',*

$$\mathbb{E}^{b_i^\diamond}\left[\frac{\partial v_i}{\partial \theta_i}\left(d\left(\theta_i', \theta_{-i}\right), \theta_i, \theta_{-i}\right) - \frac{\partial v_i}{\partial \theta_i}\left(d\left(\theta_i, \theta_{-i}\right), \theta_i, \theta_{-i}\right)\right] \cdot \left(\theta_i' - \theta_i\right) \geq 0.$$

Corollaries 2 and 3 provide single-crossing and monotonicity conditions that are 'standard' in the sense that overall they prescribe agents' marginal valuations and allocations to increase with each agent's type (either in the ex-post sense, or 'in expectation' with respect to $b^\diamond$). Compared to these, the condition in Theorem 3 is more relaxed in the sense that, if the belief restrictions admit non-trivial belief-based terms, then they may be used to 'fill' what the environment lacks in terms of the SCM conditions on the left-hand side, by relaxing the constraints on the right-hand sides of the inequality.

The belief-based terms can thus be seen as additional tools to shape agents' incentives, when standard SCM conditions are not met. The extent to which this is possible depends on the flexibility of the belief-based terms that are available to the designer, depending on the belief-restrictions. As we discussed, these are minimal in settings in which the belief sets do not vary with the type (as in belief-free settings, or in Bayesian settings with independent types, etc.), but they get larger in other cases, and more so as the belief sets get smaller.

4 Comovement of Types and Beliefs

The condition in Theorem 3 entails a certain discontinuity between settings that satisfy *generalized independence* (Def. 2), and those that do not. In the former, the only available belief-based terms are constant in m_i (cf. Corollary 6.(i) in Section 5.1.4), and hence they

[17] This Corollary generalizes known results on single-crossing and monotonicity conditions to our setting, which allows for not-everywhere differentiable allocation rules.

cannot be used to make up for failures of the SCM conditions, since the right-hand side of the condition in Theorem 3 is zero. But as soon as beliefs vary with agents' types, the possibility of using belief-based terms to recover incentive compatibility suddenly expands.

Example 3 (Ex.1, continued: Comovement of types and belief-based terms)**.** Consider the setting of Ex. 1, and replace the belief restrictions with the following formulation: $B_{\theta_i} = \{b \in \Delta(\Theta_j) : \mathbb{E}^b(\theta_j) = \gamma\frac{\theta_i}{2} + (1-\gamma)\frac{1}{2}\}$, where $\gamma \in [0,1]$ is a fixed parameter, known to the designer, that captures the degree of *comovement* between agents' beliefs and their types: for $\gamma = 1$ we obtain the baseline model from Ex. 1; for $\gamma = 0$ instead the belief restrictions satisfy *generalized independence*. Since the payoff environment is the same as in Ex. 1, ep-IC is still impossible. In fact, the canonical transfers in this setting are not $\mathcal{B}$-IC either, for any γ, and Corollary 1 and Theorem 3 jointly imply that no transfers are $\mathcal{B}$-IC when $\gamma = 0$. Now consider the following transfers:

$$t_2^{mod}(m) = t_2^*(m) - A\left(\frac{\gamma m_2^2/2 + (1-\gamma)m_2}{2} - m_1 m_2\right). \tag{7}$$

Under these belief restrictions, truthful revelation satisfies the first-order conditions, and $\frac{\partial^2 U_2^{mod}(m;\theta)}{\partial^2 m_2} = K - A\gamma/2$. Hence, $m_2 = \theta_2$ is optimal for agent 2 whenever $A > 2K/\gamma$, and hence $\mathcal{B}$-IC is possible for any $\gamma \in (0,1]$: an arbitrarily small level of *comovement* is enough to recover incentive compatibility via the design of a suitable belief-based term. □.

The insight from this example is very general, and goes beyond private values. It extends to a large class of belief restrictions, regardless of the valuation functions and of the allocation rule. The following property of the belief restrictions is key:

Definition 3. *We say that $\mathcal{B}$ admits a* responsive moment condition *if for each i there exist $L_i : \Theta_{-i} \to \mathbb{R}$ and $f_i : \Theta_i \to \mathbb{R}$ s.t. for all θ_i and $b \in B_{\theta_i}$, $\mathbb{E}^b L_i(\theta_{-i}) = f_i(\theta_i)$ where f_i is cont. diff. and f_i' is bounded away from 0.*

If, furthermore, $\mathcal{B}$ is such that, for each i and θ_i, B_{θ_i} consists of all *the beliefs $b \in \Delta(\Theta_{-i})$ such that $\mathbb{E}^b L_i(\theta_{-i}) = f_i(\theta_i)$, then we say that $\mathcal{B}$ is* maximal *with respect to the moment condition $(L_i, f_i)_{i\in I}$.*

In words, $\mathcal{B}$ admits a *moment condition* if, for every i, there exists a function of the opponents' types whose expectation given θ_i is known to the designer (i.e., for each θ_i, it is the same for all beliefs in B_{θ_i}). If such expectations are strictly monotonic in θ_i, then we say that the moment condition is *responsive*.

Moment conditions can be seen as pieces of information that the designer may have about agents' beliefs. In belief-free settings, for instance, only trivial moment conditions (where all L_i and f_i are constant) satisfy the restrictions above, and hence the designer has effectively no information about beliefs. At the oppositve extreme, in a Bayesian setting, for *any* L_i there is a f_i such that $\mathbb{E}^{b_i^\diamond} L_i(\theta_{-i}) = f_i(\theta_i)$ (albeit with $f_i' = 0$ if types are

independent, not necessarily otherwise). More broadly, the stricter the belief restrictions, the larger the set of admissible moment conditions, and hence the more information the designer has about agents' beliefs. The case when $\mathcal{B}$ is *maximal* with respect to some $(L_i, f_i)_{i\in I}$ represents the idea that the specific moment condition is essentially the *only* information about beliefs that the designer can (or is willing to) rely on.

Proposition 1. *Fix v, and let the belief restrictions admit a* responsive moment condition. *Then, for any d, there exist transfers t such that (d,t) is $\mathcal{B}$-IC.*

Hence, as long as the belief restrictions admit a responsive moment condition, then *any* allocation rule can be made $\mathcal{B}$-IC by some t. (In Example 3, $L_i(\theta_{-i}) = \theta_j$, and $f_i(\theta_i) = \frac{\gamma\theta_i+(1-\gamma)}{2}$, which satisfies the condition of the proposition if and only if $\gamma > 0$.)

The discontinuity we illustrated with Ex.3 is reminiscent of another well-known discontinuity, between Bayesian settings with *independent* and *correlated* types, namely Crémer and McLean (1985, 1988) full-surplus extraction (FSE) results.[18] But, as the next two propositions show, FSE need not obtain in these settings: Even if the designer's information about beliefs is enough to achieve the very permissive result of Proposition 1, there are bounds to the incentive compatible transfers, and information rents typically remain.

Proposition 2. *Consider a differentiable (v,d) and a $\mathcal{B}$ that is maximal with respect to a responsive moment condition $(L_i, f_i)_{i\in I}$. Then, if $(t_i)_{i\in I}$ is a $\mathcal{B}$-IC transfer scheme, for each i there exist a function $H_i : M_i \to \mathbb{R}$ such that t_i can be decomposed as follows:*

$$t_i(m) = t_i^*(m) + \int_{\underline{\theta}_i}^{m_i} \left(L_i(m_{-i}) - f_i(s)\right) H_i(s)\ ds + \tau_i(m_{-i}).$$

Moreover, there exists a continuous lower bound $K_i : \Theta_i \to \mathbb{R}$ such that, for any $\mathcal{B}$-IC transfer scheme, $\mathbb{E}^b\left[\int_{\underline{\theta}_i}^{\theta_i} \left(L_i(\theta_{-i}) - f_i(s)\right) H_i(s)\ ds\right] \geq K_i(\theta_i)$ for all θ_i and $b \in B_{\theta_i}$.

For the next proposition, we say that a function $g : \Theta \to \mathbb{R}$ is L_i-linear if it can be written in the form $g(\theta) = \delta_1(\theta_i) L_i(\theta_{-i}) + \delta_2(\theta_i)$. Also, a mechanism (d,t) is $\mathcal{B}$-*individually rational* ($\mathcal{B}$-IR) if, for each i and θ_i, $\mathbb{E}^b U_i^t(\theta_i;\theta_i) \geq 0$ for all $b \in B_{\theta_i}$.[19] Finally, we say that a mechanism *extracts the full surplus* if the individual rationality constraints hold with equality for all i, θ_i, and $b \in B_{\theta_i}$

Proposition 3. *Fix v and d, and let $\mathcal{B}$ be maximal with respect to a responsive moment condition $(L_i, f_i)_{i\in I}$. Unless for all i, $\frac{\partial v_i}{\partial \theta_i}(d(\theta), \theta)$ is L_i-linear, no transfers t can extract the full surplus.*

[18] In Bayesian settings, Crémer and McLean (1985, 1988) first studied FSE with finite types. McAfee and Reny (1992) extended the result to a continuum and to general mechanism design problems. Their condition does not always ensure *exact* FSE, but it characterizes *almost* FSE, in the sense that for any $\epsilon > 0$, there is a mechanism in which agents' surplus in the truthful equilibrium is less than ϵ. Chen and Xiong (2013) further showed that this form of FSE holds generically in the space of Bayesian models. More recent results are provided by Hu et al. (2021) and Lopomo et al. (2022), who consider distinct approaches to FSE.

[19] Recall that, for any $b \in \Delta(\Theta_{-i})$, we defined $\mathbb{E}^b U_i^t(m_i;\theta_i) := \int_{\Theta_{-i}} U_i^t(m_i, \theta_{-i}; \theta_i, \theta_{-i})\, db$. In this section we set the outside option to 0 for simplicity, but the extension to type-dependent outside options is easy.

We conclude this section with a novel sufficient condition for FSE. Our condition is stronger than McAfee and Reny's (1992), but closer in spirit to Crémer and McLean's (1988) *full rank* condition. In contrast with the work cited in footnote 18, our condition ensures an *exact* FSE result (cf. Proposition 4).Most importantly, however, and the main reason why we provide it, is that it highlights more clearly the gap between FSE and the 'responsive moment condition' above:

Definition 4. *Let $\mathcal{B}^\diamond$ be a Bayesian setting (i.e., $B^\diamond_{\theta_i} = \{b^\diamond_{\theta_i}\}$ for each i and θ_i).*

(i) We say that $\mathcal{B}^\diamond$ is differentiable *if for each i, and for any differentiable $G : \Theta \to \mathbb{R}$, the function $f_i : \Theta_i \to \mathbb{R}$, defined as $f_i(\theta_i) = \mathbb{E}^{b^\diamond_{\theta_i}}[G(\theta_i, \theta_{-i})]$, is differentiable.*

(ii) We say that $\mathcal{B}^\diamond$ satisfies the full rank condition *if, for each i, it holds that for any differentiable $g_i : \Theta_i \to \mathbb{R}$, there exists a Borel-measurable function $\kappa_i : \Theta_{-i} \to \mathbb{R}$ such that $\int_{\Theta_{-i}} \kappa_i(\theta_{-i})\, db^\diamond_{\theta_i} = g_i(\theta_i)$ for all θ_i.*

Next we show that in Bayesian settings that satisfy these conditions, not only can *any* allocation rule be made IIC, as in Proposition 1, but also the transfers can be chosen so as to match *any* target for the equilibrium expected payments:

Proposition 4. *Fix v, and let $\mathcal{B}^\diamond$ be a differentiable Bayesian setting that satisfies the full rank condition. Then, for any d and for any differentiable t, there exist transfers t' such that: (i) (d, t') is IIC; and (ii) for each i and θ_i, $\mathbb{E}^{b^\diamond_{\theta_i}}[t'_i(\theta_i, \theta_{-i})] = \mathbb{E}^{b^\diamond_{\theta_i}}[t_i(\theta_i, \theta_{-i})]$.*

These results together draw a line between the 'any d goes' result for general belief restrictions (Prop. 1), and the 'any*thing* goes' result for Bayesian settings (Prop. 4 below): while, in the latter, any interim payment functions are achievable, the extra robustness requirement in non-Bayesian settings does restrict the possible payments. We next illustrate the results of Propositions 1-4, and the key logic of their proofs, within our running example:

Example 4 (Ex. 3, continued)**.** Consider again the setting of Ex. 3, with belief restritions $B_{\theta_i} = \{b \in \Delta(\Theta_j) : \mathbb{E}^b[\theta_j] = \gamma\frac{\theta_i}{2} + (1-\gamma)\frac{1}{2}\}$. For simplicity, let us consider the case where $\gamma \in [0, 1/2]$. As we already discussed, the conditions of Prop. 1 hold, and $\mathcal{B}$-IC is attained by the transfers in eq. (7), as long as $A > 2K/\gamma$ and for any $\gamma > 0$. Figure 1 plots the range of expected payments (as a function of θ_i, for any $b \in B_{\theta_i}$) that are associated with $\mathcal{B}$-IC transfers and the condition that the lowest type pays 0.

If, however, the designer's model consists of a Bayesian setting that also satisfies the conditions of Prop. 4, then any expected payments can be induced in an incentive compatible way. For instance, let $\mathcal{B}^\diamond$ be such that, for each θ_i, $b^\diamond_{\theta_i}$ consists of a mixture of two independent uniform distributions, over $[0, \theta_i]$ and over $[0, 1]$, respectively with weights γ and $(1-\gamma)$. Then, mimicking the proof of Prop. 4, to obtain full surplus extraction we can take our 'target' transfers to be $t_i(\theta) = -v_i(d(\theta), \theta)$. With this, we can define the expected difference $g_i(\theta_i) = \int_{\Theta_j} \left(t_i - \hat{t}_i\right) db_{\theta_i}$, where $\hat{t}_i$ is a suitable IIC transfer. For agent

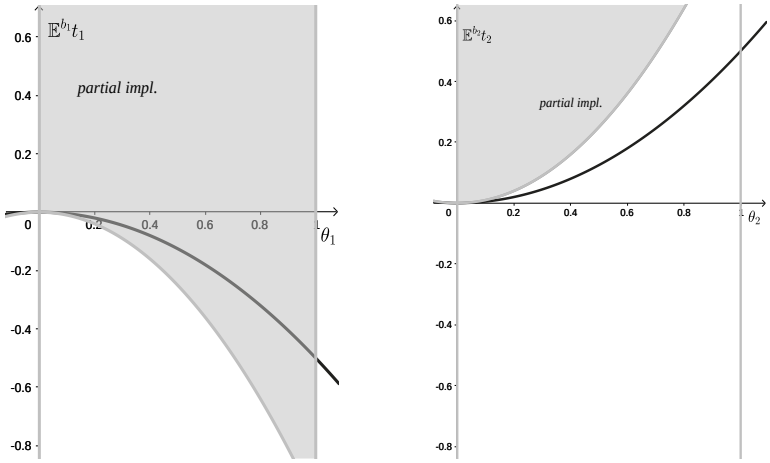

Figure 1: **Possible Expected Payments to the Agents in Ex. 3: $\mathcal{B}$-IC under $t_i(0,\theta_{-i}) \equiv 0$.** The thick black line, in both figures, is the expected canonical transfer to each agent (feasible for agent 1 but infeasible for agent 2). The gray area represents the possible interim payments under partial implementation (resulting from possibly different transfer schemes, with the restriction that the lowest type pays zero).

1, the canonical transfers are IIC, and hence they can be used in the role of $\hat{t}_1$. One has to solve the integral equation $\int_{\Theta_2} \kappa_1(\theta_2)\, db_{\theta_1} = -K\left[\gamma\frac{\theta_1^2}{2} + (1-\gamma)\frac{\theta_1}{2}\right]$ for $\kappa_1(\cdot)$. The solution is $\kappa_1(\theta_2) = \frac{K(1+\gamma)}{\gamma}\left[\theta_2(2+\gamma) + (1-\gamma)\right]$ if $\theta_2 \in [0,\gamma]$ and $\kappa_1(\theta_2) = 0$ otherwise. (See Appendix B for more on the solution of this class of integral equations.) For agent 2, we can take $\hat{t}_2(\theta) = t_2^*(\theta) - A\left(\frac{\gamma\theta_2^2/2 + (1-\gamma)\theta_2}{2} - \theta_1\theta_2\right)$ from eq. (7), which is IIC for $A > 2K/\gamma$. The integral equation $\int_{\Theta_1} \kappa_2(\theta_1)\, db_{\theta_2} = \frac{\theta_2^2}{2}\left[K(1+\gamma) - \gamma\frac{A}{2}\right] + K(1-\gamma)\frac{\theta_2}{2}$, solved for $\kappa_2(\cdot)$, yields $\kappa_2(\theta_1) = -\frac{(1-\gamma)}{\gamma}\left[\theta_1\frac{(2+\gamma)}{\gamma}\left(K(1+\gamma) - \gamma\frac{A}{2}\right) + (1-\gamma)K\right]$ if $\theta_1 \in [0,\gamma]$ and $\kappa_2(\theta_1) = 0$ otherwise. The resulting transfers, $t_i' = \hat{t}_i + \kappa_i$, preserve IIC and at the same time extract all the surplus from both agents. In fact, as per Proposition 4, any other differentiable t_i payments can be matched by constructing transfers this way. □

Hence, information rents remain, even within models where agents' beliefs might play a role in facilitating the implementation task. If the belief-restrictions are not Bayesian, even if any d can be implemented under the condition of Prop. 1, there may still be bounds to the surplus that can be extracted. The size of the information rents depends on the joint properties of the allocation rule, agents' preferences, and the belief restrictions, and they get larger as the robustness requirement strenghtens (i.e., as the belief sets get larger).

Formally, for any (v,d), and for any belief restrictions $\mathcal{B}$, let $F(\mathcal{B})$ denote the set of transfer schemes that are both $\mathcal{B}$-IC and $\mathcal{B}$-individually rational, and let $\mathcal{V}(\mathcal{B})$ denote the set of all triplets (i,θ_i,b) such that $i \in I$, $\theta_i \in \Theta_i$ and $b \in B_{\theta_i}$. Then, define:

$$\tau(\mathcal{B}) := \inf_{t \in F(\mathcal{B})} \sup_{(i,\theta_i,b)\in\mathcal{V}(\mathcal{B})} \mathbb{E}^b U_i^t(\theta_i;\theta_i)$$

if $F(\mathcal{B})$ is non-empty, and $\tau(\mathcal{B}) := \infty$ otherwise. With this, note that FSE obtains if and only if there exists $t \in F(\mathcal{B})$ such that the constraint for $\mathcal{B}$-IR holds with equality for all types of all agents, i.e. if $\tau(\mathcal{B}) = 0$. When $\tau(\mathcal{B})$ is finite but positive, in contrast, in each incentive compatible and individually rational mechanism there is at least some type that enjoys strictly positive rents. This bound to the designer's ability to extract surplus, however, decreases monotonically as belief restrictions get finer. At the extreme, if $\mathcal{B}$ is a Bayesian setting with correlated types, then FSE obtains (cf. Prop. 4). Formally:

Proposition 5. *For any (v, d), and for any $\mathcal{B}$: $\mathcal{B}' \subseteq \mathcal{B}$ implies $\tau(\mathcal{B}') \leq \tau(\mathcal{B})$. Moreover, if $\tau(\mathcal{B}^{BF}) > 0$, then there exist $\mathcal{B}$ and $\mathcal{B}'$ such that:*[20] *(i) $\mathcal{B}$ admits a responsive moment condition (Def. 3) and is such that $0 < \tau(\mathcal{B}) < \infty$; (ii) $\mathcal{B}' \subset \mathcal{B}$ and is such that $\tau(\mathcal{B}') = 0$.*

The weak monotonicity of $\tau(\cdot)$ with respect to set inclusion follows directly from the definition of $\mathcal{B}$-IC. The rest of the proposition states that – unless the environment is trivial – there always exist belief restrictions $\mathcal{B}$ in which FSE is not possible, despite $\mathcal{B}$ already granting maximal flexibility in implementing any allocation rule via belief-based terms. FSE can be achieved, but only by relying on extra information $\mathcal{B}' \subset \mathcal{B}$ about beliefs. Hence, in essentially any environment beliefs can play a meaningful role to expand the possibility of implementation, without entailing FSE.

5 Other Results, Observations and Applications

5.1 Further Implications of Theorem 1

Theorem 1 implies that identifying the set of belief-based terms is crucial to understand the limits of incentive compatibility. For some belief-restrictions, identifying this set, or some of its key properties, is relatively straightforward and delivers immediately interesting insights on the incentive compatible transfers. We discuss a few cases:

5.1.1 Belief-Free Settings

In *belief-free* settings, $\mathcal{B}^{BF}$, the condition in (5) is required to hold for all beliefs about Θ_{-i}, including degenerate ones, which is only possible if β_i is constant in m_i. Hence, a transfer scheme is $\mathcal{B}^{BF}$-IC (that is, ep-IC) only if it coincides with the canonical transfers, up to a function that is constant in agents' own reports. Thus, when all beliefs are allowed, there are no non-trivial belief-based terms. In this sense, the classical result in point (i) of Lemma 1 obtains as a special case of Theorem 1:

Corollary 4. *If t is $\mathcal{B}^{BF}$-IC, then, $\forall i$, $\beta_i(m) := t_i(m) - t_i^*(m)$ is constant in m_i.*

[20] Note that $\tau(\mathcal{B}^{BF}) = 0$ only holds in trivial environments, in which each v_i is constant in own type.

5.1.2 Bayesian Settings

In a *Bayesian setting*, $\mathcal{B}^\diamond$, for any agent i and for any function $G_i : M \to \mathbb{R}$ that is Lebesgue-integrable with respect to m_i, the term $f_i(\theta_i) := \mathbb{E}^{b^\diamond_{\theta_i}} G_i(\theta_i, \theta_{-i})$ is uniquely pinned down by the collection $(b^\diamond_{\theta_i})_{\theta_i \in \Theta_i}$ of agent i's beliefs. Hence, letting

$$\beta_i(m) := \int_{\underline{\theta}_i}^{m_i} G_i(s, m_{-i})\, ds - \int_{\underline{\theta}_i}^{m_i} f_i(s)\, ds,$$

we obtain a belief-based term, since β_i thus defined satisfies the condition in eq. (5).

In this sense, Bayesian settings are maximal in the set of belief-based terms they admit, since they can be generated starting from any arbitrary $G_i : M \to \mathbb{R}$. This is in stark contrast with the belief-free case, which as seen admits no non-trivial belief-based terms, and hence essentially no incentive compatible transfers other than the canonical ones. Here, the richness of belief-based terms gives rise to a multitude of IIC transfers, which may be used to attain different objectives beyond incentive compatibility. The results in the previous section showed how this richness, and the associated freedom to choose such functions, can be used to obtain full-surplus extraction. Other results in the literature have also exploited this richness, to pursue for instance budget balance, surplus extraction, supermodularity, contractiveness, or uniqueness (see references in footnote 7). By identifying the key condition on the belief-based terms, Theorem 1 unifies these results and lays the ground to a systematic understanding of the possibilities, and particularly the limits, of IIC.

5.1.3 Independent Types

In Bayesian settings with independent types, the belief sets not only are all singletons, but also contain the same distribution for all types of a player: for each i, $\mathcal{B}^\diamond_{\theta_i} = \{b^\diamond_i\}$ for all $\theta_i \in \Theta_i$. Then, the condition in eq. (5) implies that, for any belief-based term, its expected value at the truthful profile is constant in the agent's own type. This is stated formally in point 1 of the next Corollary. In turn, it also implies the following two points:

Corollary 5. *Let $\mathcal{B}^\diamond$ be a Bayesian environment with independent types, and let $b^\diamond_i \in \Delta(\Theta_{-i})$ denote agent i's beliefs, regardless of his type. Then:*
(i) If t_i is $\mathcal{B}^\diamond$-IC, then $\exists \kappa_i \in \mathbb{R}$ s.t. $\mathbb{E}^{b^\diamond_i} \beta_i(m_i, \theta_{-i}) = \kappa_i$ for all m_i.
(ii) If t_i is $\mathcal{B}^\diamond$-IC, $\exists \kappa_i \in \mathbb{R}$ s.t. $\mathbb{E}^{b^\diamond_i} t_i(\theta_i, \theta_{-i}) = \mathbb{E}^{b^\diamond_i} t^*_i(\theta_i, \theta_{-i}) + \kappa_i$ for all θ_i.
(iii) (d, t) is $\mathcal{B}^\diamond$-IC for some t if and only if (d, t^*) is $\mathcal{B}^\diamond$-IC.

Point (ii) is Myerson's (1981) *revenue equivalence*, here stated for general environments with interdependent values and independently distributed types. Point (iii) says that an allocation rule is partially implementable in *interim* (or *Bayes-Nash*) *equilibrium* if and only if it is implemented by t^*. But even with independence, and notwithstanding the payoff-equivalence of all IIC transfers, there may still be a value in characterizing the full set, if the designer has other objectives beyond mere incentive compatibility.

Example 5 (Independence and Multiplicity)**.** Consider two agents, with sets of types $\Theta_i = [0,1]$ and valuation functions $v_i(x,\theta) = (\theta_i + \gamma\theta_j)\,x$, for each i and $j \neq i$, where $x \geq 0$ denotes the quantity of a public good, and γ is a parameter of preference interdependence. With cost of production $c(x) = x^2/2$, the efficient allocation is $d(\theta) = (1+\gamma)(\theta_1 + \theta_2)$. In this case, the canonical transfers coincide with the generalized VCG mechanism.

Assume that types are i.i.d. draws from the uniform distribution over $[0,1]$. Then, Corollary 5 implies that IIC is possible if and only if the VCG transfers are IIC. In turn, Corollary 3 ensures that this is the case if and only if $\gamma \geq -1$. Nox let $\gamma = 3/2$, and consider the following transfers:

$$t_i^{full} = t_i^{VCG} + \alpha_i \left(m_j - \frac{1}{2}\right)(1+\gamma)m_i$$

With $\gamma = 3/2$, the VCG transfers are IIC. Furthermore, since $\mathbb{E}^b[\theta_j|\theta_i] = 1/2$ for all θ_i, these transfers satisfy both conditions in Theorem 2 for any α_i. While this richness of transfers is redundant from the viewpoint of IIC alone, it may still be useful for other purposes. For instance, if one also cares about unique implementation, it can be shown that with $\gamma = 3/2$ truthful revelation is the only rationalizable strategy if and only if $\alpha_i \in (1/2, 5/2)$. In fact, for $\alpha_i = \gamma$, truthful revelation is an *interim* dominant strategy (Ollár and Penta, 2017). □

5.1.4 On *Weak* Generalized Independence, Robust Revenue Equivalence, and the Equivalence between $\mathcal{B}$-IC and ep-IC

The logic above points to another interesting implication of Theorem 1, which suggests introducing the following weakening of *generalized independence*:

Definition 5. *$\mathcal{B}$ satisfies* **weak *generalized independence*** *if $\bigcap_{\theta_i \in \Theta_i} B_{\theta_i} \neq \emptyset$ for all $i \in I$.*

This is clearly weaker than *generalized independence* (Def. 2), which in turn encompasses as special cases both *belief-free* settings and Bayesian models with *independent types*. In these environments, Theorem 1 implies the following:

Corollary 6. *Let $\mathcal{B}$ satisfy* weak generalized indepence*, and let $p_i \in \cap_{\theta_i \in \Theta_i} B_{\theta_i}$. Then:*
(i) For any belief-based term $\beta_i : M \to \mathbb{R}$, $\exists \kappa_i \in \mathbb{R}$ s.t. $\mathbb{E}^{p_i}\beta_i(m_i, \theta_{-i}) = \kappa_i$ for all m_i.
(ii) If t_i is $\mathcal{B}$-IC, then $\exists \kappa_i \in \mathbb{R}$ s.t. $\mathbb{E}^{p_i} t_i(\theta_i, \theta_{-i}) = \mathbb{E}^{p_i} t_i^*(\theta_i, \theta_{-i}) + \kappa_i$ for all θ_i.

The discussion that follows Corollary 5 applies to any belief that belong to the intersection of the belief sets. Point (i) in Corollary 6, in particular, extends revenue equivalence to the beliefs in the intersection in such non-Bayesian settings as well. Clearly, since Def. 5 is a weakening of Def. 2, this result applies under *generalized independence* too, where also the statement of Corollary 1 holds true.

These results can be further extended to shed light on some influential results in Lopomo et al. (2021) and in Jehiel et al. (2012)). To this end, we say that $\mathcal{B}$ satisfies 'full dimensionalilty and local independence' if for each i and θ_i (i) B_{θ_i} is full dimensional (i.e., its

continuous measures span the space of continuous functions) and (ii) for each θ_i and for all $p \in B_{\theta_i}$, there exists a neighborhood of θ_i, $\mathcal{N}_{\theta_i} \subseteq \Theta_i$, such that $p \in B_{\theta_i'}$ for all $\theta_i' \in \mathcal{N}_{\theta_i}$. Then, Theorem 1 implies the following (the proof is in the Appendix):

Corollary 7. *Let $\mathcal{B}$ satisfy 'full dimensionality and local independence.' Then:*
(i) If t_i is $\mathcal{B}$-IC, then $\exists \kappa_i \in \mathbb{R}^{M_{-i}}$ s.t. $t_i(m) = t_i^*(m) + \kappa_i(m_{-i})$ for all m.
(ii) There exists a $\mathcal{B}$-IC t_i if and only if t_i^* is $\mathcal{B}$-IC.

Lopomo et al. (2021) showed that, under standard single-crossing and monotonicity assumptions, the 'full dimensionality' condition on the overlap of the belief sets implies that there is no gap between the possibility of $\mathcal{B}$-IC and ep-IC. Jehiel et al. (2012), however, gave an example of local robust implementability in a non-standard single-crossing environment, where ep-IC is impossible. Corollary 7 clarifies how close these robustness requirements are to requiring ep-IC: Under 'full dimensionality and local independence', $\mathcal{B}$-IC is possible if and only if it is achieved by the canonical transfers. Under standard ep-SCM conditions, the canonical transfers are ep-IC (Corollary 2), and hence our results also imply that there is no gap between the possibility of ep-IC and $\mathcal{B}$-IC.But without ep-SCC, as in our general setting, the canonical transfers may be $\mathcal{B}$-IC without necessarily being ep-IC.[21] Hence, without single-crossing and monotonicity, $\mathcal{B}$-IC and ep-IC need not coincide, while *revenue equivalence* still holds (Corollary 7.(i)).

5.2 Equilibrium Payoffs: An Envelope Formulation

Theorem 3 implies the following characterization of the equilibrium payoffs under $\mathcal{B}$-IC:

Theorem 4 (Payoff Characterization)**.** *Fix belief restrictions $\mathcal{B}$ and allocation rule d. For each i, let $D_i \subseteq \mathbb{R}^M$ denote the set of all belief-based terms that satisfy the inequalities of Theorem 3. Then, $(U_i)_{i\in I} \in \times_{i\in I}\mathbb{R}^{\Theta}$ is a feasible payoff-function in the truthful equilibrium of a $\mathcal{B}$-IC mechanism if and only if, for each i, there exists $\beta_i \in D_i$ such that*

$$U_i(\theta_i, \theta_{-i}; \theta) = \int_{\underline{\theta_i}}^{\theta_i} \frac{\partial v_i}{\partial \theta_i}(d(s, \theta_{-i}), s, \theta_{-i})\, ds + \beta_i(\theta_i, \theta_{-i}). \tag{8}$$

This formulation of the equilibrium payoffs resembles well-known envelope conditions that characterize the equilibrium payoffs of incentive compatible transfers. In fact, Theorem 4 generalizes several such results along different dimensions. It also highlights the limitations of pursuing an envelope approach either when beliefs do not fall within certain special cases, or when the designer has other objectives beyond mere incentive compatibility.

To see this, first suppose that the environment is *belief-free*. Then, by Corollary 4, the set D_i only contains $\beta_i : M \to \mathbb{R}$ that are constant in m_i, and hence (8) boils down to the standard envelope condition (3) in Milgrom and Segal (2002). More generally, for belief-restrictions that satisfy *generalized independence* (cf. Def. 2), and letting $b \in \cap_{\theta_i\in\Theta_i} B_{\theta_i}$,

[21] Ollár and Penta (2023) provide an example of this possibility within the context of the $\mathcal{B}^{id}$-restrictions.

then all $\beta_i \in D_i$ are such that $\mathbb{E}^b(\beta_i)$ is constant in m_i (Corollary 1), and thus the formula in (8) again delivers the standard condition for the interim expected payoffs, $\mathbb{E}^b(U_i)$, here generalized to accommodate both the possibility of interdependent values as well as non-Bayesian settings with *generalized independence*.

Thus, when $\mathbb{E}^b(\beta_i)$ is constant in m_i for all $\beta_i \in D_i$, the interim expected equilibrium payoffs under incentive compatibility are effectively pinned down, up to a constant in own message, and hence the formula gives the incentive compatible transfers as well, by using the fact that $U_i(m, \theta) = v_i(d(m), \theta) + t_i(m)$. But when the set D_i is richer, there may be a non-trivial multiplicity of payoff functions, each with its own envelope condition. In these cases (e.g., in Bayesian settings with correlated types), the payoff function is only determined once the transfers are fixed, and hence the envelope formula cannot be used to recover the incentive compatible transfers. The multiplicity of transfers determines a family of envelopes, one for each distinct belief-dependent term in D_i.

Finally, even when the envelope approach can be used to obtain transfers that are incentive compatible whenever possible (as under generalized independence), it still overlooks the richness of the set of incentive compatible transfers, which may be useful for other purposes beyond incentive compatibility. For instance, in Bayesian settings with independent types, the expected payments for all IIC transfers only differ up to a constant in own message. Such transfers, however, may induce different payoffs at non-equilibrium profiles, and hence exhibit different properties with respect to other objectives, such as uniqueness, budget balance, etc. (cf. Example 5). In this sense, also in such settings the envelope approach is more limited than the first-order approach that we pursue in this paper.

5.3 Responsive Moment Conditions and Unique Implementation

In this Section we use the characterization of the belief-based terms in Theorems 1-3, jointly with the sufficient condition of Ollár and Penta (2017), to derive easy-to-check possibility results for unique implementation under comovement.

Unique implementation requires truthful implementation to be the only strategy consistent with players' common belief in rationality and in the $\mathcal{B}$-restrictions. For each θ_i, we let $R_i^{\mathcal{B}}(\theta_i)$ denote the set of $\mathcal{B}$-rationalizable messages, that correspond to these assumptions.[22]

Definition 6 (Unique Implementation)**.** *The transfer scheme $t = (t_i)_{i \in I}$ uniquely $\mathcal{B}$-implements d if $R_i^{\mathcal{B}}(\theta_i) = \{\theta_i\}$ for all θ_i and all i.*

The next sufficient conditions for unique $\mathcal{B}$-implementation follow directly from Theorem 2 above, and from Theorem 1 in Ollár and Penta (2017):

[22] Formally, for every i and θ_i, the set of conjectures that are consistent with $\mathcal{B}$ is defined as $C_{\theta_i} := \left\{\mu_i \in \Delta(M_{-i} \times \Theta_{-i}) : marg_{\Theta_{-i}} \mu_i \in B_{\theta_i}\right\}$. Then, given a transfer scheme t, for each agent $i \in I$ let $R_i^{\mathcal{B},0} = \Theta_i \times M_i$ and for each $k = 1, 2, ...$, define the sets $R_{-i}^{\mathcal{B},k-1} = \times_{j \neq i} R_j^{\mathcal{B},k-1}$, $R_i^{\mathcal{B},k} = \{(\theta_i, m_i) : m_i \in BR_{\theta_i}(\mu_i)$ for some $\mu_i \in C_{\theta_i} \cap \Delta\left(R_{-i}^{\mathcal{B},k-1}\right)\}$, and $R_i^{\mathcal{B}} = \bigcap_{k \geq 0} R_i^{\mathcal{B},k}$. Then, $R_i^{\mathcal{B}}(\theta_i) := \{m_i : (\theta_i, m_i) \in R_i^{\mathcal{B}}\}$.

Proposition 6. *Consider a twice differentiable (v, d), and let $\mathcal{B}$ admit a responsive moment condition $(L_i, f_i)_{i \in I}$ in which each f_i is strictly increasing, $f_i' > 0$. Then, unique $\mathcal{B}$-implementation via transfers is possible if any of the following conditions holds:*

1. Highly Sensitive Moments: $\sum_{j \neq i} |\partial_j L_i (m_{-i})| < f_i' (m_i)$
2. Symmetric Canonical Substitutes: *For all i and $j \neq i$, $\partial^2_{ij} U_i^* \equiv \gamma_i < 0$, $\partial^2_{ii} U_i^* < 0$, and $\partial_j L_i (m) \equiv l_i \in \mathbb{R}_{++}$.*
3. Symmetric Canonical Complements: *for all i and $j \neq i$, $\partial^2_{ij} U_i^* \equiv \gamma_i > 0$, $\partial^2_{ii} U_i^* < 0$ and $\partial_j L_i (m) \equiv l_i \in \mathbb{R}_{++}$, and $f_i'/l_i < |\partial^2_{ii} U_i^*|/\gamma_i$. (In this case, truthful revelation is a dominant strategy).*

The condition in point 1 states that the moment condition is strongly responsive to each agent i's own type, compared to how the L_i function depends on each of the opponents' types. The conditions in points 2 and 3 refer to settings where the canonical transfers are ep-IC (condition $\partial^2_{ii} U_i^* < 0$), and that induce, respectively, symmetric strategic substitutability (condition $\partial^2_{ij} U_i^* \equiv \gamma_i < 0$ for all i and $j \neq i$) and complementarity (condition $\partial^2_{ij} U_i^* \equiv \gamma_i > 0$ for all i and $j \neq i$).[23] In the first case, unique implementation obtains if the moment conditions take the form $L_i(\theta_{-i}) = l_i \sum_{j \neq i} \theta_j$, for some $l_i > 0$. In the second case, if this holds with $f_i'/l_i < |\partial^2_{ii} U_i^*|/\gamma_i$, which in fact ensures dominant-strategy implementation.

In our running example, we illustrated how the general results in Theorems 1-3 can be used to bound the surplus that can be extracted under $\mathcal{B}$-incentive compatibility. Similarly, the characterization of the $\mathcal{B}$-IC transfers can also be used to explore the limits of unique implementation. For instance, consider the class of 'symmetric quadratic public-good environments' (SQPG) that are analyzed in Bergemann and Morris (2009a): n agents, with types $\theta_i \in [0, 1]$, valuation functions $v_i(x, \theta) = (\theta_i + \gamma \sum_{j \neq i} \theta_j) x$, where x denotes the quantity of public good. With quadratic production costs, the efficient allocation rule is $d^*(\theta) = (1 + (n-1)\gamma) \sum_{i=1,\ldots,n} \theta_i$.

In these settings, the canonical transfers (i.e., the VCG) are ep-IC if and only if $\gamma \geq -1/(n-1)$, and they achieve belief-free unique implementation if and only if $|\gamma(n-1)| < 1$. If $\gamma < -1/(n-1)$, the ep-SCM conditions fail and U_i^* are convex; if $|\gamma(n-1)| > 1$, strategic externalities are too strong, and uniqueness fails. In each of the two cases, belief-based terms can be designed to 'fix' the problem, using the results in Proposition 1 or in Proposition 6, respectively. But suppose that the two issues are present at the same time, for instance if $(n-1)\gamma < -1$. Then, Theorems 1-3 above, jointly with Theorem 1 in Ollár and Penta (2025), imply that even for belief restrictions that satisfy the sufficient condition of both Propositions 1 and 6, unique $\mathcal{B}$-implementation is impossible:

[23] Strategic substitutes, for instance, arise naturally in efficient implementation problems with public goods, where agents' marginal utility for the public good is increasing in others' types.

Proposition 7. *In a SQPG environment with* $(n-1)\gamma < -1$*, if* $\mathcal{B}$ *is maximal with respect to a responsive moment condition* $(L_i, f_i)_{i\in I}$ *such that* $f_i' > 0$ *for all* i *and* $\partial_j L_i > 0$ *for all* $j \neq i$*, then unique* $\mathcal{B}$*-implementation is impossible.*

6 Related Literature

This paper contributes to the literature on robust mechanism design, particularly following the approach in Bergemann and Morris (2005), that is to achieve implementation of a given allocation rule for a large set of beliefs. The first wave of this literature focused on *belief-free* environments. More specifically, Bergemann and Morris (2005, 2009a,b) study belief-free implementation in static settings, respectively in the partial, full and virtual implementation sense. The belief-free approach has been extended to dynamic settings by Müller (2016) and Penta (2015). Penta (2015) considers environments in which agents may obtain information over time, and applies a dynamic version of rationalizability based on a backward induction logic (cf. Penta (2011) and Catonini and Penta (2022)). Müller (2016) instead studies virtual implementation via dynamic mechanisms, in a static belief-free environment, using a stronger version of rationalizability with forward induction.

Belief restrictions, as a general framework to accommodate varying degrees of robustness, was first introduced by Ollár and Penta (2017) to study how beliefs can be used to attain *unique* implementation, and some special cases were were analyzed in Ollár and Penta (2022, 2023, 2025), in settings where incentive compatibility followed directly from standard assumptions. In this paper, in contrast, we focused on the more fundamental question of how beliefs can be used via the design of the transfers for the very establishment of incentive compatibility. An earlier instance of belief restrictions can also be found in Artemov et al. (2013), which focuses instead on virtual implementation in environments where the baseline belief-free approach of Bergemann and Morris (2009b) is enriched with a collection of (commonly known) first-order beliefs. More recently, belief restrictions have also proven useful within the area of *behavioral mechanism design*, to model features of individuals' beliefs that cannot be cast within the standard framework (see, e.g., Gagnon-Bartsch, Pagnozzi, and Rosato (2021), and Gagnon-Bartsch and Rosato (2024)).

From a methodological viewpoint, we pursued a generalization of the classical *first-order approach* that identifies necessary conditions for *local* incentive compatibility constraints (cf. Rogerson, 1985; Jewitt, 1988), and then studies sufficient conditions for global optimality. Carvajal and Ely (2013) also studied the design of incentive compatible mechanisms in settings in which the envelope formula cannot be used, due to non-convexity or non-differentiability of the valuations, but only within standard Bayesian settings.

A few papers have used special cases of belief restrictions to model robustness with respect to *local* perturbations around a given Bayesian belief-setting. For instance, Jehiel et al. (2012) show that, under certain restrictions on preferences, minimal notions of ro-

bustness are as demanding as the belief-free case. A similar result is proven in Lopomo et al. (2021), for overlapping beliefs, and in Lopomo et al. (2022), within an auction setting. As discussed, these results are in line with those we obtain under generalized independence (cf. Corollary 1). The exact connections between our results and those of these papers are discussed in Section 5. In terms of the framework, the belief-restrictions that we consider encompass the belief sets studied by the above papers. In contrast to those papers, we develop a first-order approach and also provide several possibility results for transfer design under various degrees of robustness. Lopomo et al. (2021), on the other hand, also consider more general preferences, which are beyond the scope of our work. Their notion of overlapping beliefs, however, excludes the belief restrictions that enable our possibility results and characterizations under *comovement* (Propositions 1, 2, 3, and 5 in Section 4).

Several alternative approaches to robustness have been put forward, which we view as complementary. For instance, Börgers and Smith (2012, 2014), focus on the role of eliciting beliefs to weakly implement a correspondence in a belief-free setting. Börgers and Li (2019) provide a more systematic analysis of implementation relying on first-order beliefs. Other approaches model robustness with respect to certain behavioral concerns directly in the implementation concept. These include criteria such as credibility of the designer (Akbarpour and Li (2020)), a behavioral notion of strong strategy proofness (Li (2017)), safety considerations with respect to model misspecification (Gavan and Penta (2023)), convergence of best response dynamics (Mathevet (2010); Mathevet and Taneva (2013); Healy and Mathevet (2012), and Sandholm (2002, 2005, 2007)), etc. Yet another approach is based on maxmin criteria, as pursued for example by Chung and Ely (2007); Chassang (2013); Carroll (2015); Yamashita (2015); He and Li (2022). The aim here is typically to explore whether 'natural' mechanisms can be justified as worst-case optimal, within a suitable robustness set (see Carroll (2019) for a survey of this literature). In this paper, in contrast, we fix an allocation rule and require implementation not only for the worst-case beliefs, but for all beliefs in the robustness set. In this sense, our approach is closest to the original approach of Bergemann and Morris (2005, 2009a,b).

7 Conclusions

We studied incentive compatibility in a general framework for *robust mechanism design* that can accommodate various degrees of robustness, including both belief-free (e.g., Bergemann and Morris (2005, 2009a,b)) and standard Bayesian settings as special cases, as well as accommodate interesting settings for *behavioral mechanism design*, to model features of individuals' beliefs that cannot be cast within the standard framework (e.g., Gagnon-Bartsch, Pagnozzi, and Rosato (2021), and Gagnon-Bartsch and Rosato (2024)).

For general *belief restrictions*, we characterized the set of incentive compatible direct mechanisms in general environments with interdependent values. The necessary conditions

that we identified provide a unified view of several known results, as well as novel ones, including a *robust* version of the *revenue equivalence* theorem that holds under a notion of *generalized independence* that also applies to non-Bayesian settings. Our general sufficient conditions imply that, under weak properties on the belief restrictions, it is possible to achieve implementation even in environments that violate standard single-crossing and monotonicity conditions, and we provide an explicit design for the implementing transfers.

From a methodological perspective, we showed that, in spite of its simplicity, a suitable generalization of the classical *first-order approach* (e.g., Laffont and Maskin, 1980; Rogerson, 1985; Jewitt, 1988, etc.), allows a wealth of novel results: (i) on the one hand, it identifies the class of incentive compatible transfers in settings which cannot be handled with the standard envelope approach (such as in Bayesian settings with correlated types, or with general (non-Bayesian) belief restrictions); (ii) on the other hand, even in settings where the equilibrium payoffs are pinned down by the envelope approach, it identifies the richness of incentive compatible transfers that may serve purposes beyond incentive compatibility.[24]

More broadly, our main results develop a general design principle, centered around the design of *belief-based terms*, in pursuit of various objectives in mechanism design. We showed that minimal information about agents' beliefs may suffice to implement *any* allocation rule. Yet, if the setting is non-Bayesian, information rents are generally possible, and they get larger the less information the designer has about agents' beliefs. Our *belief restrictions* may thus capture a meaningful notion of 'comovement' of beliefs and types that is useful for implementation, but without incurring into the pitfalls of 'full-surplus extraction' results (cf. Crémer and McLean, 1985, 1988). This framework may thus favor mechanism design's reappropriation of environments with non-exclusive information, in which distilling intuitive and reliable economic intuition has long appeared elusive, within the prevailing paradigm.

References

Akbarpour, M. and S. Li (2020). Credible auctions: A trilemma. *Econometrica 88*(2), 425–467. 27

Artemov, G., T. Kunimoto, and R. Serrano (2013). Robust virtual implementation: Toward a reinterpretation of the wilson doctrine. *Journal of Economic Theory 148, no. 2*, 424–447. 26

Bergemann, D. and S. Morris (2005). Robust mechanism design. *Econometrica*, 1771–1813. 2, 6, 26, 27

[24] Some examples of extra desiderata in the literature, which our approach enables to unify, include budget balance (d'Aspremont and Gérard-Varet, 1979), surplus extraction (Crémer and McLean, 1985, 1988), stability (Mathevet, 2010; Mathevet and Taneva, 2013; Healy and Mathevet, 2012; Sandholm, 2002, 2005, 2007), and uniqueness (Ollár and Penta, 2017, 2022, 2023), etc.).

Bergemann, D. and S. Morris (2009a). Robust implementation in direct mechanisms. *The Review of Economic Studies 76*(4), 1175–1204. 2, 6, 25, 26, 27

Bergemann, D. and S. Morris (2009b). Robust virtual implementation. *Theoretical Economics 4*(1), 45–88. 2, 6, 26, 27

Bewley, T. (2002). Knightian decision theory: Part i. *Decisions in Economics and Finance 25*, 79–110. 2, 8

Börgers, T. and J. Li (2019). Strategically simple mechanisms. *Econometrica 87*(6), 2003–2035. 27

Börgers, T. and D. Smith (2012). Robustly ranking mechanisms. *American Economic Review 102*(3), 325–329. 27

Börgers, T. and D. Smith (2014). Robust mechanism design and dominant strategy voting rules. *Theoretical Economics 9*(2), 339–360. 27

Carroll, G. (2015). Robustness and linear contracts. *American Economic Review 105*(2), 536–563. 27

Carroll, G. (2019). Robustness in mechanism design and contracting. *Annual Review of Economics 11*, 139–166. 27

Carvajal, J. C. and J. C. Ely (2013). Mechanism design without revenue equivalence. *Journal of Economic Theory 148*, 104–133. 2, 26

Catonini, E. and A. Penta (2022). Backward induction reasoning beyond backward induction. *TSE Working Paper*. 26

Chassang, S. (2013). Calibrated incentive contracts. *Econometrica 81*(5), 1935–1971. 27

Che, Y. and J. Kim (2006). Robustly collusion-proof implementation. *Econometrica 74*, 1063–1107. 5

Chen, Y.-C. and S. Xiong (2013). Genericity and robustness of full surplus extraction results. *Econometrica 81*(1), 825–847. 17

Chung, K.-S. and J. C. Ely (2007). Foundations of dominant-strategy mechanisms. *The Review of Economic Studies 74*(2), 447–476. 27

Crémer, J. and R. P. McLean (1985). Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica 53*(2), 345–361. 1, 4, 5, 17, 28

Crémer, J. and R. P. McLean (1988). Full extraction of the surplus in bayesian and dominant strategy auctions. *Econometrica*, 1247–1257. 1, 2, 4, 5, 17, 18, 28

Dasgupta, P. and E. Maskin (2000). Efficient auctions. *The Quarterly Journal of Economics 115*(2), 341–388. 8

d'Aspremont, C. and L.-A. Gérard-Varet (1979). Incentives and incomplete information. *Journal of Public Economics 11*(1), 25–45. 5, 28

Ennuschat, L. P. and A. Penta (2025). Partial implementation and belief restricitons. Technical report. 2, 8

Gagnon-Bartsch, T., M. Pagnozzi, and A. Rosato (2021). Projection of private values in auctions. *American Economic Review 111*(10), 3256–3298. 2, 26, 27

Gagnon-Bartsch, T. and A. Rosato (2024). Quality is in the eye of the beholder: taste projection in markets with observational learning. *American Economic Review 114*(11), 3746–3787. 2, 26, 27

Gavan, M. J. and A. Penta (2023). Safe implementation. *BSE working paper*. 27

Green, J. and J.-J. Laffont (1977). Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica*, 427–438. 8

He, W. and J. Li (2022). Correlation-robust auction design. *Journal of Economic Theory 200*, 105403. 27

Healy, P. J. and L. Mathevet (2012). Designing stable mechanisms for economic environments. *Theoretical economics 7*(3), 609–661. 5, 27, 28

Hochstadt, H. (1989). *Integral equations*. Wiley Classics Library. John Wiley & Sons. 38

Hu, N., J. Haghpanah, and R. Hartline (2021). Full surplus extraction from samples. *Journal of Economic Theory 193*. 17

Jehiel, P. and L. Lamy (2018). A mechanism design approach to the tiebout hypothesis. *Journal of Political Economy 126*(2), 735–760. 8

Jehiel, P., M. Meyer-ter Vehn, and B. Moldovanu (2012). Locally robust implementation and its limits. *Journal of Economic Theory 147*(6), 2439–2452. 22, 23, 26

Jehiel, P., M. Meyer-ter Vehn, B. Moldovanu, and W. R. Zame (2006). The limits of ex post implementation. *Econometrica 74*(3), 585–610. 6

Jehiel, P. and B. Moldovanu (2001). Efficient design with interdependent valuations. *Econometrica 69*(5), 1237–1259. 6, 8

Jewitt, I. (1988). Justifying the first-order approach to principal-agent probblems. *Econometrica*, 1177–1190. 2, 11, 26, 28

Laffont, J.-J. and D. Martimort (1997). Collusion under asymmetric information. *Econometrica*, 875–911. 5

Laffont, J.-J. and E. Maskin (1980). A differential approach to dominant strategy mechanisms. *Econometrica*, 1507–1520. 8, 28

Li, S. (2017). Obviously strategy-proof mechanisms. *American Economic Review 107*(11), 3257–3287. 27

Lopomo, G., L. Rigotti, and C. Shannon (2021). Uncertainty in mechanism design. *arXiv:2108.12633*. 22, 23, 27

Lopomo, G., L. Rigotti, and C. Shannon (2022). Uncertainty and robustness of surplus extraction. *Journal of Economic Theory 199*, 105088. 17, 27

Mathevet, L. (2010). Supermodular mechanism design. *Theoretical Economics 5*(3), 403–443. 5, 27, 28

Mathevet, L. and I. Taneva (2013). Finite supermodular design with interdependent valuations. *Games and Economic Behavior 82*, 327–349. 5, 27, 28

McAfee, R. P. and P. J. Reny (1992). Correlated information and mecanism design. *Econometrica*, 395–421. 17, 18

Milgrom, P. R. (2004). *Putting auction theory to work*. Cambridge University Press. 8

Milgrom, P. R. and I. Segal (2002). Envelope theorems for arbitrary choice sets. *Econometrica 70*(2), 583–601. 23

Müller, C. (2016). Robust virtual implementation under common strong belief in rationality. *Journal of Economic Theory 162*, 407–450. 6, 26

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research 6*(1), 58–73. 4, 8, 21

Neeman, Z. (2004). The relevance of private information in mechanism design. *Journal of Economic Theory 117*, 55–77. 2

Ollár, M. (2024). Incentive compatibility with multi-dimensional types: the role of belief restricitons. Technical report. 6

Ollár, M. and A. Penta (2017). Full implementation and belief restrictions. *American Economic Review 107*(8), 2243–2277. 2, 5, 6, 7, 22, 24, 26, 28, 38

Ollár, M. and A. Penta (2022). Efficient full implementation via transfers: Uniqueness and sensitivity in symmetric environments. Volume 112, pp. 438–443. American Economic Association. 5, 26, 28

Ollár, M. and A. Penta (2023). A network solution to robust implementation: The case of identical but unknown distributions. *Review of Economic Studies 90*(5), 2517–2554. 5, 7, 12, 14, 23, 26, 28, 38

Ollár, M. and A. Penta (2025). Robust unique implementation. Technical report. 14, 25, 26, 38

Penta, A. (2011). Backward induction reasoning in games with incomplete information. *University of Winconsin-Madison*. 26

Penta, A. (2015). Robust dynamic implementation. *Journal of Economic Theory 160*, 280–316. 6, 26

Postlewaite, A. and D. Schmeidler (1986). Implementation in differential information economies. *Journal of Economic Theory 39*, 14–33. 4

Rogerson, W. P. (1985). The first-order approach to principal-agent probblems. *Econometrica*, 1357–1367. 2, 11, 26, 28

Safronov, M. (2018). Coalition-proof full efficient implementation. *Journal of Economic Theory 177*, 659–677. 5

Sandholm, W. H. (2002). Evolutionary implementation and congestion pricing. *The Review of Economic Studies 69*(3), 667–689. 5, 27, 28

Sandholm, W. H. (2005). Negative externalities and evolutionary implementation. *The Review of Economic Studies 72*(3), 885–915. 5, 27, 28

Sandholm, W. H. (2007). Pigouvian pricing and stochastic evolutionary implementation. *Journal of Economic Theory 132*(1), 367–382. 5, 27, 28

Segal, I. (2003). Optimal pricing mechanisms with unknown demand. *American Economic Review 93*(3), 509–529. 8

Wilson, R. (1987). *Game-theoretic analyses of trading processes.* Advances in Economic Theory. in Bewley (ed.), Cambridge University Press. 2

Yamashita, T. (2015). Implementation in weakly undominated strategies: Optimality of second-price auction and posted-price mechanism. *The Review of Economic Studies 82*(3), 1223–1246. 27

Appendix

A Proofs

Proof of Theorem 1. Fix an agent i. First, we show that $t_i^*(m)$ is well-defined since the allocation rule d is p.diff.[25] Since v_i is twice continuously differentiable, $\frac{\partial v_i}{\partial \theta_i}$ is continuously differentiable over $X \times \Theta$. Now, for fixed m_{-i}, $\frac{\partial v_i}{\partial \theta_i}(d(\cdot, m_{-i}), \cdot, m_{-i})$ – a function from M_i to $\mathbb{R}$ – is a composite function of d and $\frac{\partial v_i}{\partial \theta_i}$ and since d is piecewise differentiable over Θ_i, we have that for all m_{-i}, $\frac{\partial v_i}{\partial \theta_i}(d(\cdot, m_{-i}), \cdot, m_{-i})$, a function from M_i to $\mathbb{R}$, is piecewise continuous, therefore integrable, over M_i.

Claim 1: t_i^* is p.diff over M.

Proof of Claim 1: Recall that $t_i^*(m) = -v_i(d(m), m) + \int_{\underline{\theta_i}}^{m_i} \frac{\partial v_i}{\partial \theta_i}(d(s, m_{-i}), s, m_{-i})\, ds$. Since d is p.diff, restricted to its pieces, $\frac{\partial v_i}{\partial \theta_i}(d(\cdot), \cdot) : M \to \mathbb{R}$ is continuously differentiable over the same pieces as v_i is twice cont.diff. Therefore $\int^{m_i} \frac{\partial v_i}{\partial \theta_i}$ is p.diff over M, and thus t_i^* is p.diff over M.

Now, consider a piecewise differentiable $\mathcal{B}$-IC t_i, and we let $\beta_i := t_i - t_i^*$. Then, by Claim 1, β_i is p.diff over M. Next, since t_i is $\mathcal{B}$-IC, for all θ_i, $b \in B_{\theta_i}$, we have that, when the derivative exists, $\left[\partial_i \mathbb{E}^b\big(v_i(d(m_i, \theta_{-i}), \theta) + t_i(m_i, \theta_{-i})\big)\right]\big|_{m_i=\theta_i} = 0$. Since the canonical transfer t^* by its construction satisfies the ex-post FOC, the above statement holds for t_i^* too. Now, from this, for $t_i - t_i^*$, for all θ_i and $b \in B_{\theta_i}$ for which both derivatives exist, we have $\left[\partial_i \mathbb{E}^b\big(t_i - t_i^*\big)(m_i)\right]\big|_{m_i=\theta_i} = 0$. Next, we use the following claim to extend this result to all differentiability points of $\mathbb{E}^b \beta_i$, beyond the joint differenttiability points of $\mathbb{E}^b t_i$ and $\mathbb{E}^b t_i^*$. □

Claim 2: For a p.diff $f : M \to \mathbb{R}$ and $b \in \Delta(\Theta_{-i})$ with p.diff cdf, $\mathbb{E}^b f : M_i \to \mathbb{R}$ is p.diff.

Proof of Claim 2: Consider b's cdf. which has finitely many pieces: $S_1^b, \ldots, S_K^b$. Write $\mathbb{E}^b f(m_i) = \int_{\Theta_{-i}} f(m_i, \theta_{-i})\, db = \sum_{j=1}^K \int_{int\ S_j^b} f(m_i, \theta_{-i})\, db$. For each j, let $A_j(m_i) := \int_{int\ S_j^b} f(m_i, \theta_{-i})\, db$. Since f is p.diff over M, it is p.diff over each S_j^b and it has finitely many pieces of S_j^b: $S_{j,1}^b, \ldots, S_{j,l}^b, \ldots, S_{j,L_j}^b$. Rewrite A_j such that $A_j(m_i) = \sum_{l=1}^{L_j} \int_{int\ S_{j,l}^b} f(m_i, \theta_{-i})\, db$, and note that f is continuouse over $int\ S_{jl}^b$. Therefore $A_j : M_i \to \mathbb{R}$ is p.diff over M_i for each j. Since $\mathbb{E}^b f$ is a sum of K such functions, it is p.diff over M_i (that is, it has at most finitely many jumps). □

Note that by Claim 2, if b has a p.diff cdf, then $\mathbb{E}^b v_i$ is p.diff and thus $\mathbb{E}^b t_i^*$ is p.diff, which also means that $\mathbb{E}^b(t_i - t_i^*)$ is p.diff, moreover, it is differentiable in the joint differentiability points of $\mathbb{E}^b t_i$ and $\mathbb{E}^b t_i^*$, that is, over M_i with the exception of at most finitely many points. Therefore, if $\mathbb{E}^b \beta_i(\cdot)$ has further differentiability points, then the expected value condition must extend to these as well, and hence the Theorem follows. ■

Remark. As this is clear from the last part of the proof above, for a belief $b \in B_{\theta_i}$ which has a p.diff cdf,[26] $\mathbb{E}^b \beta_i$ is almost everywhere differentiable on M_i. Thus the expected

[25] For example, consider two agents. The single item allocation rule given by the allocation probabilities $d_1(\theta) = 1 - d_2(\theta) = \{1 \text{ if } \theta_1 > \theta_2; 1/2 \text{ if } \theta_1 = \theta_2; 0 \text{ otherwise}\}$ satisfies our definition of piecewise differentiability.

[26] Note that for example, discrete distributions, full support continuous distributions, as well as their

value condition of Theorem 1, for typically considered belief-restrictions, implies substantial restrictions on what form the function β_i can take.

Proof of Corollary 1. It follows from Corollary 6. ■

Proof of Theorem 2. By the assumed differentiability, β_i is also twice continuously differentiable and as the domains are compact, by the Leibniz rule, (1) obtains from Theorem 1. Further, under t_i, reporting θ_i is locally optimal and thus (2) obtains from the decomposition of the payoff function into U_i^* and β_i. In the other direction, if (2) holds strictly for all m_i, then the expected payoff function is strictly concave, and by the decomposition and (1), the FOC holds at θ_i, hence t_i is $\mathcal{B}$-IC. ■

Characterization of Belief-based Terms in Example 1.

Claim: Consider the belief-restrictions $\mathcal{B}^\gamma$; for all $i \in \{1,2\}$ and for all θ_i, $B^\gamma_{\theta_i} = \left\{b \in \Delta(\Theta_j) : \mathbb{E}^b \theta_j = \gamma_i \theta_i\right\}$. In the special case of $\gamma_i = 1/2$, this is the setting considered in Ex. 1. Recall that $\theta_i \in [0,1]$ and we assume that $0 < \gamma_i < 1$. Then a function $\beta_i : M \to \mathbb{R}$ which is differentiable in m_i is a belief-based term if and only if for some real functions H_i on M and τ_i on M_{-i}, it takes the form $\beta_i(m) = \int_0^{m_i} \left(s - \frac{m_j}{\gamma_i}\right) H_i(s)\, ds + \tau_i(m_{-i})$.

Proof of the Claim. First, if β_i is of the given form, then $\partial_i \beta_i(m_i, m_j) = \left(m_i - \frac{m_j}{\gamma_i}\right) H_i(m_i)$ which for all θ_i, at the truthtelling profile for all beliefs in B_{θ_i} satisfies the expected value condition, thus it is a belief-based term. Second, in the other direction, if β_i is a differentiable belief-based term, then by the point-beliefs in $B^\gamma_{\theta_i}$, we have that (i) $\partial_i \beta_i(\theta_i, \gamma_i \theta_i) = 0$ for all θ_i. Next, we show that $\partial_i \beta_i : M \to \mathbb{R}$ is linear in m_j. This is so, as $B^\gamma_{\theta_i}$ contains beliefs that place non-zero probabilities on two points x and y which give a splitting of $\gamma_i \theta_i$: there is a probability α such that $\alpha x + (1-\alpha) y = \gamma_i \theta_i$. Note that such α exists for any points that are such that $x \leq \gamma_i \theta_i \leq y$. Each of these beliefs imply, by the expected value condition, that $\alpha \partial_i \beta_i(\theta_i, x) + (1-\alpha)\, \partial_i \beta_i(\theta_i, y) = 0$ as well. Hence for any fixed m_i, $\partial_i \beta_i$ is linear in m_j. Hence, there are functions f_1 and f_2 onM_i for which $\partial_i \beta_i(m) = f_1(m_i)\, m_j + f_2(m_i)$. At the same time, as by (i) above, these functions must be such that for all θ_i, $f_1(\theta_i)\, \gamma_i \theta_i + f_2(\theta_i) = 0$. From this and by change of notation for the functions, $\beta_i(m)$ has the form as claimed. Finally, the initial condition of "0 type pays 0" of this example implies that $\tau_i \equiv 0$ and so β_i takes the form as stated in Ex. 1. □

Remark. Notice that this result is a consequence of linear algebraic principles. Consider a closed set X and a continuous function $f : X \to \mathbb{R}$ from the usual normed vector space of the continuous functions with the L_1 norm. The set of continuous functions in $\mathcal{M} := \{\mu \in \Delta(X) : \mathbb{E}^\mu f = 0\}$ spans the orthogonal complement of f and thus a function g from the same space satisfies $\mathbb{E}^\mu g = 0$ for all $\mu \in \mathcal{M}$ only if $g = \alpha f$ for some $\alpha \in \mathbb{R}$. Applying this in the setting of the above example, for fixed θ_i: X is Θ_{-i}, f is $\theta_j - \theta_i/\gamma_i$ and g is $\partial_i \beta_i$ and it must satisfy the expected value condition in Eq. 5. Thus for any fixed θ_i, $\partial_i \beta_i$ must be in the linear space of $\theta_j - \theta_i/\gamma_i$, that is for some $\alpha_{\theta_i} \in \mathbb{R}$, $\partial_i \beta_i = \alpha_{\theta_i}(\theta_j - \theta_i/\gamma_i)$. We generalize this characterization of belief-based terms by using this observation later in the

convex combinations have piecewise differentiable cdfs and are Borel-measures.

Proofs of Prop. 2 and 3.

Proof of Theorem 3. The payoffs $U_i = v_i + t_i^* + \beta_i$, by using (1) and adding and subtracting $\int_{m_i}^{\theta_i} \frac{\partial v_i}{\partial \theta_i}\left(d\left(s, m_{-i}\right) s, m_{-i}\right) ds + \beta_i\left(\theta_i, m_{-i}\right)$, can be rewritten, at the profile $m_{-i} = \theta_{-i}$, as

$$U_i\left(m_i, \theta_{-i}; \theta\right) = \int_{\underline{\theta_i}}^{\theta_i} \frac{\partial v_i}{\partial \theta_i}\left(d\left(s, \theta_{-i}\right), s, \theta_{-i}\right)\ ds + \beta_i(\theta)$$
$$- \int_{m_i}^{\theta_i} \underbrace{\left(\frac{\partial v_i}{\partial \theta_i}\left(d\left(s, \theta_{-i}\right), s, \theta_{-i}\right) - \frac{\partial v_i}{\partial \theta_i}\left(d\left(m_i, \theta_{-i}\right), s, \theta_{-i}\right)\right)}_{=: \mathcal{SC}_i\left(m_i, s, \theta_{-i}\right)} ds + \beta_i\left(m_i, \theta_{-i}\right) - \beta_i(\theta).$$

The first two terms do not depend on the report m_i, and the latter three terms give 0 if $m_i = \theta_i$. Thus $m_i = \theta_i$ is best response if and only if the expected gain from misreport, $-\mathbb{E}^b \int_{m_i}^{\theta_i} \mathcal{SC}_i\left(m_i, s, \theta_{-i}\right) ds + \mathbb{E}^b \beta_i\left(m_i\right) - \mathbb{E}^b \beta_i\left(\theta_i\right)$, is nonpositive; which is the condition from the inequality of this theorem. ■

Proof of Proposition 1. For each agent i, let $t_i := t_i^* - A_i\left(\int^{m_i} f_i(s)\, ds - L_i\left(m_{-i}\right) m_i\right)$. By the smoothness and implied boundedness assumptions on v and d, the left-hand side of the inequality in Theorem 3 is bounded, and hence there exists A_i large (resp., small) enough if f_i is increasing (resp., decreasing) such that the inequality in Theorem 3 holds for $\beta_i(m) = -A_i\left(\int^{m_i} f_i(s)\, ds - L_i\left(m_{-i}\right) m_i\right)$. ■

Proof of Proposition 2. Fix agent i. If $\mathcal{B}$ is maximal with respect to $(L_i, f_i)_{i \in I}$, then any belief-based term β_i satisfies the necessary condition of Theorem 1 if and only if $\partial_i \beta_i = \left(L_i\left(m_{-i}\right) - f_i\left(m_i\right)\right) H_i\left(m_i\right)$, where H_i is a real function over M_i.[27] Then, if t_i is $\mathcal{B}$-IC, by Theorem 1, it can be written as,

$$t_i(m) = t_i^*(m) + \int_{\underline{\theta_i}}^{m_i} \left(L_i\left(m_{-i}\right) - f_i(s)\right) H_i(s)\ ds + \tau_i\left(m_{-i}\right).$$

Next, we need to check when the SOC at the truthful profile holds.[28] To this end, we need to study when it is the case that for all $b_{\theta_i} \in B_{\theta_i}$,

$$\partial_{ii}^2 \mathbb{E}^{b_{\theta_i}} U_i^*\left(m_i, \theta_{-i}, \theta\right)\Big|_{m_i = \theta_i} + \partial_{ii}^2 \mathbb{E}^{b_{\theta_i}} \beta_i\left(m_i, \theta_{-i}\right)\Big|_{m_i = \theta_i} \leq 0$$

$$-\mathbb{E}^{b_{\theta_i}}\left(\frac{\partial^2 v_i\left(d(\theta), \theta\right)}{\partial x \partial \theta_i} \frac{\partial d(\theta)}{\partial \theta_i}\right) \leq f_i'\left(\theta_i\right) H_i\left(\theta_i\right)$$

Let us set

$$\overline{SCM}_i\left(\theta_i\right) := \sup_{b_{\theta_i} \in B_{\theta_i}} \mathbb{E}^{b_{\theta_i}}\left(-\frac{\partial^2 v_i\left(d(\theta), \theta\right)}{\partial x \partial \theta_i} \frac{\partial d(\theta)}{\partial \theta_i}\right).$$

[27] One direction is clear. To prove the other direction, recall the following. Consider a closed set X and a continuous function $f : X \to \mathbb{R}$ from the usual normed vector space of continuous functions with the L_1 norm. The set of continuous functions in $\mathcal{M} := \{\mu \in \Delta(X) : \mathbb{E}^\mu f = 0\}$ spans the orthogonal complement of f and thus a function g from the same space satisfies $\mathbb{E}^\mu g = 0$ for all $\mu \in \mathcal{M}$ only if $g = \alpha f$ for some $\alpha \in \mathbb{R}$. Applying this here, for fixed θ_i: X is Θ_{-i}, f is $L_i\left(\theta_{-i}\right) - f_i\left(\theta_i\right)$ and g is $\partial_i \beta_i$ and it must satisfy the expected value condition in Eq. 5. Thus for any fixed θ_i, $\partial_i \beta_i$ must be in the linear space of $L_i\left(\theta_{-i}\right) - f_i\left(\theta_i\right)$.

[28] The canonical externalities are $\partial_{ij}^2 U_i^*(m, \theta) = \left(\frac{\partial^2 v_i(\theta, d(m))}{\partial^2 x} \frac{\partial d}{\partial \theta_j} - \frac{\partial^2 v_i(m, d(m))}{\partial x \partial \theta_j} - \frac{\partial^2 v_i(m, d(m))}{\partial^2 x} \frac{\partial d}{\partial \theta_j}\right) \frac{\partial d}{\partial \theta_i} + \left(\frac{\partial v_i(\theta, d(m))}{\partial x} - \frac{\partial v_i(m, d(m))}{\partial x}\right) \frac{\partial^2 d}{\partial \theta_j \partial \theta_i}$.

With this notation, if $f_i' > 0$, then $\overline{SCM}_i/f_i'$ is a lower bound on H_i and if $f_i' < 0$, then $\overline{SCM}_i/f_i'$ is an upper bound on H_i. Next, consider the modification of the interim payments and notice that the order of integration can be exchanged:

$$\mathbb{E}^{b_{\theta_i}}\beta_i(\theta) = \mathbb{E}^{b_{\theta_i}}\int_{\underline{\theta_i}}^{\theta_i}\left(L_i(\theta_{-i}) - f_i(s)\right)H_i(s)\ ds$$

$$= \int_{\underline{\theta_i}}^{\theta_i}\left(\mathbb{E}^{b_{\theta_i}}L_i(\theta_{-i}) - f_i(s)\right)H_i(s)\ ds = \int_{\underline{\theta_i}}^{\theta_i}\left(f_i(\theta_i) - f_i(s)\right)H_i(s)\ ds.$$

First, if $f_i' > 0$, then the weights on H_i are positive, and the lower bound on H_i gives a lower bound on the second term. Therefore $\mathbb{E}^{b_{\theta_i}}\beta_i(\theta) \geq \int_{\underline{\theta_i}}^{\theta_i}\left(f_i(\theta_i) - f_i(s)\right)\left[\overline{SCM}_i/f_i'\right](s)\ ds$. Second, if $f_i' < 0$, then the upper bound on H_i gives a lower bound on the second term, hence, in this case too, the same inequality holds. ■

Proof of Proposition 3. By way of contradiction, assume that t is $\mathcal{B}$-IC and extracts the surplus. By Theorem 1, t_i can be written as $t_i(m) = t_i^*(m) + \int_{\underline{\theta_i}}^{m_i}\left(L_i(m_{-i}) - f_i(s)\right)H_i(s)\ ds + \tau_i(m_{-i})$. Moreover, for all θ_i and $b \in B_{\theta_i}$, $\mathbb{E}^b U_i^t(\theta;\theta) = 0$. Using the formula in 1, and the calculation for $\mathbb{E}^{b_{\theta_i}}\int_{\underline{\theta_i}}^{\theta_i}\left(L_i(\theta_{-i}) - f_i(s)\right)H_i(s)\ ds = \int_{\underline{\theta_i}}^{\theta_i}\left(f_i(\theta_i) - f_i(s)\right)H_i(s)\ ds$ as in the Proof of Prop. 2, these impy that

$$\mathbb{E}^b\left(\int_{\underline{\theta_i}}^{\theta_i}\frac{\partial v_i}{\partial\theta_i}\left(d(s,\theta_{-i})s,\theta_{-i}\right)\ ds + \tau_i(\theta_{-i})\right) = -\int_{\underline{\theta_i}}^{\theta_i}\left(f_i(\theta_i) - f_i(s)\right)H_i(s)\ ds.$$

The RHS of this expression depends on θ_i but not on b, therefore the LHS must be the same for all $b \in B_{\theta_i}$. Applying the argument of Footnote 27 to these functions, we have that the function $\int_{\underline{\theta_i}}^{\theta_i}\frac{\partial v_i}{\partial\theta_i}\left(d(s,\theta_{-i})s,\theta_{-i}\right)\ ds + \tau_i(\theta_{-i})$ must be L_i-linear. This function is differentiable in θ_i and thus its derivative $\frac{\partial v_i}{\partial\theta_i}\left(d(\theta),\theta\right)$ must be L_i-linear as well. In summary, unless $\frac{\partial v_i}{\partial\theta_i}\left(d(\theta),\theta\right)$ is L_i-linear, $\mathcal{B}$-IC and FSE lead to a contradiction. ■

Proof of Proposition 4. First note that if $\mathcal{B}^\diamond$ is differentiable and satisfies the full rank condition, then there exist functions $(L_i, f_i)_{i\in I}$ that satisfy the condition of Prop. 1. Then, for each i, consider $\hat{t}_i := t_i^* - A_i\left(\int^{m_i} f_i(s)\,ds - L_i(m_{-i})\,m_i\right)$. From the proof of Prop. 1, $(d,\hat{t})$ is IIC for A_i large (small) enough if f_i is increasing (decreasing). Next, let $g_i : \Theta_i \to \mathbb{R}$ be defined as $g_i(\theta_i) := \int_{\Theta_{-i}}[t_i(\theta_i,s) - \hat{t}_i(\theta_i,s)]db^\diamond_{\theta_i}$ and note that, by construction and Def. 4, g_i is differentiable in θ_i. Using the full rank condition, let $\kappa_i : \Theta_{-i} \to \mathbb{R}$ be s.t. $\int_{\Theta_{-i}}\kappa_i(\theta_{-i})db^\diamond_{\theta_i} = g_i(\theta_i)$ for each θ_i. Then, letting t_i' be defined as $t_i'(\theta_i,\theta_{-i}) := \hat{t}_i(\theta_i,\theta_{-i}) + \kappa_i(\theta_{-i})$, the direct mechanism (d,t') is both IIC and such that $\mathbb{E}^{b^\diamond_{\theta_i}}[t_i'(\theta_i,\theta_{-i})] = \mathbb{E}^{b^\diamond_{\theta_i}}[t_i(\theta_i,\theta_{-i})]$. ■

Proof of Proposition 5. Fix (v,d). The first inequality follows from the relaxed robustness requirement. The rest of the proposition requires the construction of the two belief-restrictions $\mathcal{B}$ and $\mathcal{B}'$. Note that for each i, there is a function $L_i : M_{-i} \to \mathbb{R}$ such that $\frac{\partial v_i}{\partial\theta_i}\left(d(\theta),\theta\right)$ is not L_i-linear. For each i fix $\gamma_i \in (0,1)$, and let the belief-restrictions $\mathcal{B}$ be maximal with respect to the responsive moment condition $(L_i, \gamma_i\theta_i)_{i\in I}$.Prop. 1 implies that

$\mathcal{B}$-IC transfers exist, thus $F(\mathcal{B})$ is non-empty and $\infty > \tau(\mathcal{B})$. Yet, as a consequence of Prop. 3, FSE is not possible, that is, $\tau(\mathcal{B}) > 0$. Next, let $\mathcal{B}'$ be s.t. $B'_{\theta_i} = \{p_{\theta_i}\}$ and s.t. (i) p_{θ_i} has a pdf that is continuouse and non-zero over the support $\times_{j\neq i}\left[\underline{\theta}_j, \underline{\theta}_j + (\theta_i - \underline{\theta}_i)(l_j/l_i)\right]$, where for each i, $l_i := \overline{\theta}_i - \underline{\theta}_i$, and (ii) for all θ_i, $\mathbb{E}^{p_{\theta_i}} L_i(\theta_{-i}) = \gamma_i \theta_i$. (Note that for each θ_i, matching the fixed first moment is possible.) For $\mathcal{B}'$ thus constructed, the construction in Ex. 3 shows that a t exists which ensured FSE and is $\mathcal{B}$-IC and hence $\mathcal{B}'$-IC as well. ■

Proof of Corollary 4. By Theorem 1, for every $b \in \Delta(\Theta_{-i})$, at each point of differentiability, $\partial_i \mathbb{E}^b \beta_i(m_i, \theta_{-i}) = 0$. In particular, this holds for all point-beliefs, and thus for all fixed m_{-i}, in all points of differentiability of $\beta_i(\cdot, m_{-i})$, we have $\partial_i \beta_i(m_i, \theta_{-i}) = 0$. Thus for each fixed m_{-i}, the function $\beta_i(\cdot, m_{-i})$ can jump at most finitely many times, and on its pieces, the derivative is 0, therefore on its pieces, it must be constant. However, if it had a jumping point, then by the smoothness properties of v_i, it would violate incentive compatibility. Therefore β_i must be constant everywhere in m_i. ■

Proof of Corollary 5. Let $\mathcal{B}^\diamond$ be a Bayesian environment with independent types. Note that by independence the belief does not change with the type, so let $b_i^\diamond \in \Delta(\Theta_{-i})$ denote agent i's beliefs, regardless of his type. First, recall that $\mathbb{E}^{b_i^\diamond}[\beta_i(\cdot, \theta_{-i})]$ is a function over M_i that can jump at most finitely many times. In its points of differentiability, the derivative is 0, thus the function is constant. If the function would jump, it would violate incentive compatibility, hence it is a constant κ_i over M_i, which proves (1) of this corollary. By the characterization in Theorem 1, (2) and (3) follow. ■

Proof of Corollary 6. The proof of Corollary 5 applies to belief $p_i \in \cap_{\theta_i \in \Theta_i} \Delta(\Theta_{-i})$. ■

Proof of Corollary 7. First, we present the proof for the case when t_i is differentiable. Fix θ_i. By Theorem 1, $\beta_i := t_i - t_i^*$ satisfies the expected value condition, that is for each $p \in B_{\theta_i}$, $\mathbb{E}^p \partial_i \beta_i = 0$ and thus, since B_{θ_i} is full dimensional, $\partial_i \beta_i$ must be constant in θ_{-i}, therefore $\mathbb{E}^p \partial_i \beta_i = \partial_i \beta_i = 0$. Consider a neighborhood $\mathcal{N}_{\theta_i}$ s.t. $p \in B_{\theta'_i}$ for all $\theta'_i \in \mathcal{N}_{\theta_i}$. By the previous, $\partial_i \beta_i$ equals a function over Θ_i that is 0 everywhere on $\mathcal{N}_{\theta_i}$. Applying the same argument to each θ_i and by the continuity of β_i, we have that $\exists \kappa_i \in \mathbb{R}^M$ s.t. $\beta_i(m) = \kappa_i(m_{-i})$.

Next, when t_i is only p.diff., then the argument can be applied to the pieces of differentiability of the corresponding functions and by incentive compatibility, one can show that β_i can not jump in m_i. ■

Proof of Theorem 4. Consider the payoff equation of the Proof of Theorem 3. By setting $m_i = \theta_i$, the theorem follows. ■

Proof of Proposition 6. By the characterization of $\mathcal{B}$-IC transfers for responsive moment conditions in the Proof of Prop. 2, we have that if $\mathcal{B}$ admits the moment condition $(L_i, f_i)_{i\in I}$, then the transfers given by constant $H_i(m_i) \equiv h_i$ such that $t_i = t_i^* + h_i(L_i(m_{-i}) - f_i(s))$ satisfy the first-order conditions, moreover imply the following second-order properties: for

all i and $j \neq i$, and for all (m, θ)

$$\partial^2_{ii} U_i(m, \theta) = \partial^2_{ii} U^*_i(m, \theta) - h_i \cdot f'_i(m_i),$$

$$\partial^2_{ij} U_i(m, \theta) = \partial^2_{ij} U^*_i(m, \theta) + h_i \cdot \partial_j L_i(m_{-i}).$$

Hence, the sufficient condition for $\mathcal{B}$-IC and uniquness from Ollár and Penta (2017) implies that if for all i

$$\max_{(\theta,m)} \left(\partial^2_{ii} U^*_i(m, \theta) - h_i \cdot f'_i(m_i) \right) < 0 \text{ and}$$

$$\max_{(\theta,m)} \sum_{j \neq i} |\partial^2_{ij} U^*_i(m, \theta) + h_i \cdot \partial_j L_i(m_{-i})| < \min_{(\theta,m)} |\partial^2_{ii} U^*_i(m, \theta) - h_i \cdot f'_i(m_i)|,$$

then full $\mathcal{B}$-Implementation follows. Next, (1) under *highly sensitive moments*, setting h_i to a negative number with a high enough magnitude, or (2) under *symmetric canonical substitutes*, setting h_i to γ_i / l_i, or (3) under *symmetric canonical complements*, setting h_i to γ_i / l_i ensures that both inequalities hold and Full $\mathcal{B}$-Implementation attains. ■

Proof of Proposition 7. Fix a $\mathcal{B}$-IC t and note that such a t exists by Prop. 1. Note that if $(n-1)\gamma < -1$, then for all i, $\partial^2_{ii} U^*_i > 0$, and thus by the characterization of $\mathcal{B}$-IC transfers under responsive moment condition in Prop. 2, following the notation from there, we must have $H_i(m_i) > 0$. But this implies with regards to strategic externalitites, under the given assumption on the responsive moments, that $\partial^2_{ij} U^t_i = \partial^2_{ij} U^*_i + \partial_j L_i(m_{-i}) H_i(m_i) > \partial^2_{ij} U^*_i (= -\gamma > 0)$. Thus by the assumption on γ, for t, the corresponding strategic externality matrix is such that its largest absolute eigenvalue is larger than 1, which by Lemma 1 (ii) in Ollár and Penta (2023) (or by its generalization, Theorem 1 of Ollár and Penta, 2025) implies that Full $\mathcal{B}$-Implementation fails. ■

B On Example 3: Beliefs and the Inverse Problem

Consider an agent with type θ_i and beliefs given such that $\theta_j | \theta_i = \gamma \nu_{\theta_i} + (1 - \gamma) \eta_{ij}$ where ν_{θ_i} is $U[0, \theta_i]$ and, independently of this, η_{ij} is $U[0, 1]$. Let us examine the solvability of $\int_0^1 \alpha_i(\theta_j) p(\theta_j | \theta_i) d\theta_j = f(\theta_i)$. (For a thorough mathematical treatment on the solvability of integral equations we recommend the book Hochstadt (1989).) The pdf of the conditional random variable is such that:

if $1 - \gamma > \gamma \theta_i$,

$$p(\theta_j | \theta_i) = \begin{cases} \frac{1}{\gamma \theta_i (1-\gamma)} \theta_j & \text{if } \theta_j \in (0, \gamma \theta_i) \\ \frac{1}{1-\gamma} & \text{if } \theta_j \in [\gamma \theta_i, 1 - \gamma) \\ \frac{1 - \gamma + \gamma \theta_i - \theta_j}{\gamma \theta_i (1-\gamma)} & \text{if } \theta_j \in [1 - \gamma, 1 - \gamma + \gamma \theta_i) \\ 0 & \text{otherwise} \end{cases}$$

and if $1 - \gamma < \gamma \theta_i$

$$p\left(\theta_j|\theta_i\right) = \begin{cases} \frac{1}{(1-\gamma)\gamma\theta_i}\theta_j & \text{if } \theta_j \in (0, 1-\gamma) \\ \frac{1}{\gamma\theta_i} & \text{if } \theta_j \in [1-\gamma, \gamma\theta_i) \\ \frac{1-\gamma+\gamma\theta_i-\theta_j}{(1-\gamma)\gamma\theta_i} & \text{if } \theta_j \in [\gamma\theta_i, 1-\gamma+\gamma\theta_i) \\ 0 & \text{otherwise} \end{cases}.$$

There are two cases to be considered: either $\gamma \leq 1/2$ or $\gamma > 1/2$.

Part 1: If $\gamma \leq 1/2$, then for all θ_i, $1-\gamma > \gamma\theta_i$. Let us look for solutions of the form such that $\alpha_i\left(\theta_j\right)$ is 0 outside of $\theta_j \in [0, \gamma]$. In this case, since $\theta_i < \frac{1-\gamma}{\gamma}$ for all θ_i, $\int_0^1 \alpha_i\left(\theta_j\right) p\left(\theta_j|\theta_i\right) d\theta_j = f\left(\theta_i\right)$ can be written as

$$\int_0^{\gamma\theta_i} \alpha\left(\theta_j\right) \frac{\theta_j}{\left(1-\gamma\right)\gamma\theta_i}\, d\theta_j + \int_{\gamma\theta_i}^{\gamma} \alpha\left(\theta_j\right) \frac{1}{1-\gamma}\, d\theta_j = f\left(\theta_i\right).$$

Starting from this expression, in the following three lines, (1) we change variable to $s := \gamma\theta_i$ and differentiate and simplify, (2) reorganize and differentiate for a second time, (3) reorganize:

$$\begin{aligned} \int_0^s \alpha\left(\theta_j\right) \frac{-\theta_j\left(1-\gamma\right)}{\left(1-\gamma\right)^2 s^2}\, d\theta_j &= f'\left(\frac{s}{\gamma}\right)\frac{1}{\gamma} \\ \alpha\left(s\right) s &= -\left(1-\gamma\right)\left(f''\left(\frac{s}{\gamma}\right)\frac{s^2}{\gamma} + 2f'\left(\frac{s}{\gamma}\right)\frac{s}{\gamma}\right) \\ \alpha\left(s\right) &= -\left(1-\gamma\right)\left(f''\left(\frac{s}{\gamma}\right)\frac{s}{\gamma} + 2f'\left(\frac{s}{\gamma}\right)\frac{1}{\gamma}\right), \end{aligned}$$

to, finally, introduce notation $L_\gamma\left(s\right) := f''\left(\frac{s}{\gamma}\right)\frac{s}{\gamma} + 2f'\left(\frac{s}{\gamma}\right)\frac{1}{\gamma}$ and change variables to get the solution which is: for all $\theta_j \in [0, \gamma]$, $\alpha\left(\theta_j\right) = -\left(1-\gamma\right) L_\gamma\left(\theta_j\right)$, and 0 otherwise.[29]

Part 2: If $\gamma > 1/2$, then there are two cases to be considered: either $1-\gamma > \gamma\theta_i$ or $1-\gamma \leq \gamma\theta_i$. Eitherways, let us look for solutions of the form such that $\alpha_i\left(\theta_j\right)$ is 0 outside of $[\gamma, 1]$.

Case (A): $1-\gamma > \gamma\theta_i$. In this case, $\int_0^1 \alpha_i\left(\theta_j\right) p\left(\theta_j|\theta_i\right) d\theta_j = f\left(\theta_i\right)$ can be written as

$$\int_\gamma^{1-\gamma+\gamma\theta_i} \frac{1-\gamma+\gamma\theta_i-\theta_j}{\left(1-\gamma\right)\gamma\theta_i}\alpha\left(\theta_j\right)\ d\theta_j = f\left(\theta_i\right).$$

Starting from this expression, we change variable to $s := \gamma\theta_i$ and simplify and differentiate, differentiate for a second time,

$$\begin{aligned} 0 + \int_\gamma^{1-\gamma+s} \alpha\left(\theta_j\right)\ d\theta_j &= \left(1-\gamma\right)\left(f\left(\frac{s}{\gamma}\right)s\right)' \\ \alpha\left(1-\gamma+s\right) &= \left(1-\gamma\right)\left(f''\left(\frac{s}{\gamma}\right)\frac{s}{\gamma} + 2f'\left(\frac{s}{\gamma}\right)\frac{1}{\gamma}\right), \end{aligned}$$

[29] Note that $L_\gamma\left(s\right) = \left(f\left(\frac{s}{\gamma}\right)s\right)''$.

to, finally, change variables, use the notation L_γ and get the solution which is: for all $\theta_j \in [\gamma, 1]$, $\alpha(\theta_j) = (1-\gamma) L_\gamma(\theta_j - (1-\gamma))$, and 0 otherwise.

Case (B): $1 - \gamma \leq \gamma\theta_i$. In this case, $\int_0^1 \alpha_i(\theta_j) p(\theta_j|\theta_i) d\theta_j = f(\theta_i)$ can be written as

$$\int_\gamma^{\gamma\theta_i} \frac{1}{\gamma\theta_i} \alpha(\theta_j) d\theta_j + \int_{\gamma\theta_i}^{1-\gamma+\gamma\theta_i} \frac{1-\gamma+\gamma\theta_i-\theta_j}{(1-\gamma)\gamma\theta_i} \alpha(\theta_j) \ d\theta_j = f(\theta_i).$$

Starting from this expression, we change variable to $s := \gamma\theta_i$ and simplify and differentiate, differentiate for a second time,

$$\alpha(s) + 0 - \alpha(s) + \int_s^{1-\gamma+s} \frac{1}{1-\gamma} \alpha(\theta_j) d\theta_j = \left(f\left(\frac{s}{\gamma}\right) s \right)'$$

$$\alpha(1-\gamma+s) - \alpha(s) = (1-\gamma)\left(f''\left(\frac{s}{\gamma}\right)\frac{s}{\gamma} + 2f'\left(\frac{s}{\gamma}\right)\frac{1}{\gamma} \right).$$

Finally, change variables, use the notation L_γ, and the assumption on the format such that $\alpha(s)$ is 0 for all $s < \gamma$ and get the solution which is: for all $\theta_j \in [\gamma, 1]$, $\alpha(\theta_j) = 0 + (1-\gamma) L_\gamma(\theta_j - (1-\gamma))$, and 0 otherwise.

In summary, in Part 2, differentiating the integral equation twice implies a unique candidate solution since the solution suggested for Case (B) is the same as in Case (A). The candidate solution, when checked against the domain restrictions, works indeed and hence is the solution of the integral equation. □