University of Zurich[UZH]

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 188

Bilateral Trade with Loss-Averse Agents

Jean-Michel Benkert

First version: November 2014
This version: July 2022

Bilateral Trade with Loss-Averse Agents*

Jean-Michel Benkert†

This version: July 2022
First version: November 2014

Abstract

The endowment and attachment effect are empirically well-documented in bilateral trade situations. Yet, the theoretical literature has so far failed to formally identify these effects. We fill this gap by introducing expectations-based loss aversion, which can explain both effects, into the classical setting by Myerson and Satterthwaite (1983). This allows us to formally identify the endowment and attachment effect and study their impact on information rents, allowing us to show that, in contrast to other behavioral approaches to the bilateral trade problem, the impossibility of inducing materially efficient trade persists in the presence of loss aversion. We then turn to the design of optimal mechanisms and consider the problem of maximizing the designer's revenue as well as gains from trade. We find that the designer optimally provides the agents with full insurance in the money dimension and, depending on the distribution of types, optimally increases or decreases the trade frequency in the presence of loss aversion.

Keywords: Bilateral trade, loss aversion, mechanism design, endowment and attachment effect

JEL Classification: C78, D01, D02, D82, D84, D90

*This paper is a revised version of the first chapter of my PhD thesis submitted at the University of Zurich. I would like to thank Zoltán Balogh, Olivier Bochet, Juan Carlos Carbajal, Eddie Dekel, Jeff Ely, Samuel Häfner, Fabian Herweg, Heiko Karle, Botond Kőszegi, René Leal Vizcaíno, Igor Letina, Shangen Li, Shou Liu, Daniel Martin, Konrad Mierendorff, Marc Möller, Oleg Muratov, Wojciech Olszewski, Anne-Katrin Roesler, Yuval Salant, Aleksei Smirnov, Gerhard Sorger, Ran Spiegler, Egor Starkov, Tom Wilkening, Peio Zuazo Garin, and seminar participants in Bern, Zurich, at the Workshop on Mechanism Design and Behavioural Economic in Glasgow, at the ZWE 2014 and ESEM 2016 for helpful comments. I am especially grateful to my supervisors Nick Netzer and Georg Nöldeke for their guidance as well as numerous comments and suggestions. I would like to thank the University of Basel and Northwestern University for their hospitality while some of this work was conducted and the UBS International Center of Economics in Society at the University of Zurich as well as the Swiss National Science Foundation (Doc.Mobility Grant P1ZHP1_161810) for financial support. All errors are my own.

†University of Bern, Department of Economics, Schanzeneckstrasse 1, 3001 Bern, Switzerland. Email: jean-michel.benkert@unibe.ch.

1 Introduction

The bilateral trade setting describes a simple, yet economically important situation. There is a seller, who owns a good and might be willing to sell it, and a buyer, who might be interested in buying it. Different parts of the economic literature have approached this setting differently. In the field of mechanism design we assume that the agents' valuation of the good is private information and study what outcomes a designer can achieve through different institutions. Famously, Myerson and Satterthwaite (1983) have shown that some potential gains from trade will be left on the table as a rule. This impossibility result on efficiency constitutes a cornerstone within economics overall.

In the empirical literature, especially by means of experiments, we have studied how people behave in such trade situations and how the institution affects behavior. Here, two notable effects have been documented: the endowment effect, going back to Thaler (1980), and, more recently, the attachment effect (Ericson and Fuster, 2011). The endowment effect tells us that ownership of the good drives up the seller's valuation of the good. The attachment effect tells us that a buyer can get attached to a good she does not own (yet) and that this attachment drives up her valuation for it. Both of these empirical findings can be explained by the model of expectations-based loss aversion by Kőszegi and Rabin (2006, 2007), which builds on the seminal work by Kahneman and Tversky (1979).[1] In their model, essentially, people compare an outcome to some reference point, which is given by their initial, rational expectations of the outcome. In the case of the buyer, for instance, it is the expectation that she will buy the good which leads to her attachment to the good. Importantly, the neoclassical model which is employed in Myerson and Satterthwaite (1983) cannot explain why these effects would materialize. The model by Kőszegi and Rabin, however, is a natural candidate to better understand these empirical effects and their implications on trade situations from a theoretical perspective.

In this paper, we thus introduce expectations-based loss aversion into an otherwise standard mechanism-design approach to the bilateral-trade problem. More specifically, we augment the model by Myerson and Satterthwaite (1983) (henceforth MS), in which both agents have quasi-linear utility over ownership of the good and money, by allowing for both agents to have reference-dependent preferences as modeled in Kőszegi and Rabin (2006, 2007) (henceforth KR). We call the standard utility from ownership of the good and money *material utility*, and, in addition, introduce *gain-loss utility* with respect to both, money and ownership of the good, separately. The reference point, relative to which agents evaluate an outcome, is formed endogenously as the rational expectations

[1] There is a substantial empirical evidence of loss aversion, e.g., Fehr and Goette (2007), Post, van den Assem, Baltussen, and Thaler (2008), Crawford and Meng (2011) and Pope and Schweitzer (2011). In particular, see Ericson and Fuster (2014) for an excellent review on the role of loss aversion in explaining behavioral effects in exchange situations.

over the outcome.[2] We introduce the formal framework in detail in Section 2, where we also characterize incentive compatible mechanisms.

We then begin our analysis in Section 3 by identifying the theoretical counterparts of the endowment and attachment effect. In Proposition 1, we show that in any incentive compatible mechanism, loss aversion, by means of the attachment and the endowment effect, reduces the information rent of the buyer and increases the information rent of the seller, respectively. To better understand this and to fix ideas, consider the mechanism in which trade takes place whenever the buyer values the good more than the seller, i.e., whenever trade is *materially efficient*. In the absence of loss aversion, the buyer has an incentive to imitate a lower type, that is, pretend that she does not value the good as much as she actually does, in order to drive down the price she has to pay for it. The flip side of this behavior, is that by doing so, she reduces the probability of trade actually taking place. This is where expectations-based loss aversion kicks in. The possibility of getting the good induces an attachment to the good, which, if trade was to not take place, gives rise to a feeling of loss. In order to avoid this loss, which is felt more strongly than a commensurate gain, the buyer is less eager to shade her valuation than in the absence of loss aversion. Consequently, it is easier to induce truthful behavior from the buyer and thus her information rent decreases due to the attachment effect. Turning to the seller, we find that the endowment effect plays out in a similar fashion, but with the opposite result. In the absence of loss aversion, the seller wants to imitate a higher type, in order to receive a higher transfer. Loss aversion reinforces this behavior, as reporting a higher type increases the chance of trade not taking place and hence keeping the good the seller is endowed with. Thus, it becomes even harder to induce truthful behavior from the seller and her information rent increases.

The result on the effect of loss aversion on the agents' information rent in Proposition 1 is of interest for two reasons. First, as we have already noted, it formally identifies the theoretical counterparts of the attachment and endowment effect. Second, it suggests an interesting connection to Myerson and Satterthwaite's impossibility result. The standard interpretation of the impossibility result is that the gains from trade cannot cover the information rents that accrue to the agents in order to ensure incentive compatibility given the participation constraints and budget balance. Since loss aversion reduces the buyer's information rent, it could mitigate the severity of the impossibility problem or even reverse it, thus enabling the implementation of materially efficient trade. Indeed, Proposition 2 shows that the presence of a loss-averse buyer can mitigate the impossibility

[2]Ericson and Fuster (2011), Abeler, Falk, Goette, and Huffman (2011), Crawford and Meng (2011), Gill and Prowse (2012), Karle, Kirchsteiger, and Peitz (2015), and Bartling, Brandes, and Schunk (2015) provide evidence for the assumption that the reference point is determined by expectations. In contrast, see Heffetz and List (2014) and Gneezy, Goette, Sprenger, and Zimmermann (2017) for papers that show the limits of this. Heffetz (2021) provides a nuanced discussion on some of the conflicting evidence.

result in the sense that a lower subsidy would be needed to induce materially efficient trade. However, a reversal is beyond reach, as loss aversion not only reduces the buyer's information rent, but also her participation constraint becomes harder to satisfy, due to the ex-post variation in payoffs, which lower expected utility.

We would like to note that the robustness of the impossibility result in the present context is in stark contrast to other papers with non-standard preferences, which show that the impossibility result can be reversed. In the case of intentions-based social preferences the reversal is driven by the fact that the incentive compatibility constraints can be turned slack by introducing an action which generates sufficiently strong feelings of kindness, thereby essentially eliminating any tension between ex-post efficiency and the agents' incentives (Bierbrauer and Netzer, 2016). Similarly, as agents become more altruistic, their utility becomes more aligned with the expected gains from trade, reducing the tension between ex-post efficiency and the agents' incentives (Kucuksenel, 2012). Thus, in contrast to the present framework, the channel alleviating the impossibility problem does not conflict with the incentive compatibility or the participation constraints, meaning that a reversal is possible.

In Section 4 we turn to the problem of designing optimal mechanisms and begin with the problem of maximizing the designer's revenue. We show that in the presence of loss aversion any revenue-maximizing mechanism features what we call *interim-deterministic transfers*, that is, the transfer of an agent is independent of the other agent's report and is thus deterministic given her own type. This reduces ex-post variations in payoffs, thereby making loss-averse agents better off. Turning to the optimal trade rule, we note that it is not possible to simply obtain the optimal trade rule by pointwise maximization as in MS, because the agents' expected utilities endogenously depend on the mechanism through the reference point. We thus first show that the optimal trade rule must take a particular form. Namely, holding fixed the buyer's type, if trade optimally takes place for some seller type, then trade should also take place for all lower seller types. This captures the intuitively appealing notion that trade should take place for buyers with high valuations and sellers with low valuations. With this in hand we can reformulate the objective function such that pointwise maximization is once more applicable and derive the optimal mechanism. We find that the presence of loss aversion can reduce or increase the optimal amount of trade depending on the distribution of types. In some cases, for instance when types are distributed uniformly with identical support, the designer induces less trade in the presence of loss aversion. Thus, beyond eliminating all ex-post variation in the agents' transfers, thereby fully insuring them against any losses in the money dimension, the designer may partially insure agents against losses in the trade dimension by reducing the trade probability. In this case with uniformly distributed types, one can show that the designer reduces the trade probability as the stakes increase and provides the agents with

full insurance by eliminating trade altogether for sufficiently high stakes. Intuitively, as the stakes become larger, it becomes too costly to induce loss-averse agents to take on any uncertainty.

Besides maximizing the designer's revenue, another natural question to ask is to how the designer can maximize the gains arising from trade. In the presence of loss aversion one needs to clarify what the relevant welfare criterion is and how to handle gain-loss utility. The literature on behavioral welfare economics provides some guidance and allows us to distinguish between model-based and model-less approaches (Manzini and Mariotti, 2014). In a model-based approach (e.g., Benkert and Netzer, 2018; Rubinstein and Salant, 2012) the welfare criterion is developed based on an underlying theory (or, a model) of mistakes. In contrast, in a model-less approach (e.g., Apesteguia and Ballester, 2015; Bernheim and Rangel, 2009) multiple inconsistent preferences are being aggregated into a welfare criterion solely on the basis of observed choices. Thus, the designer may take different stances on how to treat gain-loss utility when aiming to maximize gains from trade. Proceeding analogously as for the revenue-maximizing mechanism, we can derive the optimal mechanism for both when the designer wants to maximize only material gains from trade or total gains from trade (including gain-loss utility). In general, the optimal mechanisms may be different for the these two distinct objectives. It turns out, however, that for the case of uniformly distributed types and symmetric degrees of loss aversion, the optimal mechanisms coincide, so that it does not matter whether the designer considers loss aversion a mistake or not.

1.1 Related literature

Most closely related to our paper is the literature on mechanism design with loss-averse agents. Eisenhuth (2019) considers the problem of a risk-neutral seller who wants to maximize revenue by selling a good to loss-averse buyers. Using the framework of KR, he finds that the optimal auction is an all-pay auction with reserve price when agents bracket narrowly. This result corresponds to our finding that transfers are interim deterministic in optimal mechanisms and, as one can show, extends beyond the auction and bilateral trade setting. Duraj (2018) considers mechanism design problems with agents who are loss averse on news utility, that is, agents' utility depends on changes in their beliefs over the outcome as in Kőszegi and Rabin (2009). In an application to bilateral trade he shows the robustness of the impossibility result in this setting.[3]

Also related is the (increasingly large) literature on behavioral industrial organization

[3]In an older version of that paper, which was made available by personal communication, Duraj showed that the impossibility result can be reversed under some conditions in the presence of news utility (Duraj, 2015). We thank Niccolò Lomys for making the connection.

with loss-averse agents.[4] Rosato (2017) considers a sequential bargaining model with a risk-neutral seller and a loss-averse buyer.[5] Also within the framework of KR, but assuming wide bracketing, he shows that the buyer's loss aversion softens the rent-efficiency trade off for the seller. As in the present paper, this is driven by the attachment effect: the buyer is willing to accept lower offers to avoid the risk of a breakdown of the negotiations.[6] In contrast to the present paper, neither Rosato (2017) (nor Eisenhuth (2019) above) feature loss-averse sellers, but only loss-averse buyers. Heidhues and Kőszegi (2014) and Rosato (2016) consider models with a monopolist selling to expectations-based loss-averse consumers. In both papers the monopolist uses random prices to induce the attachment effect, increasing the consumers willingness to pay and thus profits. In contrast, in the present paper agents are already confronted with uncertainty due to the private nature of types and there is no need to further "inject" randomness to induce the attachment or endowment effect. Indeed, the designer optimally insures agents fully against any variation in transfers and partially in the trade dimension in order to reduce ex-post variation in payoffs.

Finally, our paper also relates to the large literature on the bilateral trade problem, which has followed Myerson and Satterthwaite (1983). Arguably, the departure from the classical setting most closely related to our paper, is to consider risk-averse agents. However, in contrast to loss aversion, risk aversion cannot explain the endowment and attachment effect. Early on, Chatterjee and Samuelson (1983) showed that when agents "become infinitely risk averse" all material gains from trade can be realized using a double-auction. More recently, Garratt and Pycia (2020) examine the bilateral trade problem relaxing the assumption that the agents have quasi-linear utility.[7] Allowing for risk aversion and wealth effects, they provide conditions for the possibility of realizing all gains of trade. The impossibility result can be reversed in this setting, because the presence of risk aversion and wealth effects give rise to additional gains from trade, which then suffice to cover the agents' information rents.[8] In contrast to Garratt and Pycia (2020)

[4]See for instance Karle and Möller (2020) and the references therein.

[5]See Shalev (2002) and Driesen, Perea, and Peters (2012) for other approaches incorporating loss aversion to bargaining.

[6]The attachment effect also plays a role in a number of other papers, among others Karle and Schumacher (2017) in a model of advertisement or in Rosato (2021) who proposes expectations-based loss aversion as an explanation for the "afternoon effect" observed in sequential auctions.

[7]See also the references in Garratt and Pycia (2020) for more work on the bilateral trade problem in the classic framework with quasi-linear utility following Myerson and Satterthwaite (1983). Moreover, see Wolitzky (2016) and Crawford (2021) for analyses of the bilateral trade problem with maxmin and level-k agents, respectively.

[8]In contrast to Garratt and Pycia (2020), we obtain quasi-linear utility due to narrow-bracketing of gain-loss utility and having piece-wise linear value functions. Thus, the relaxation of quasi-linear utility, which gives rise to the possibility result in their paper, is not present in our framework. The narrow-bracketing assumption also sets the present setting apart from that in Gershkov, Moldovanu, Strack, and Zhang (2021), who study optimal auction design when agents have constant relative risk aversion. They find that agents' utility is, in the language of the present paper, interim deterministic, i.e., does

we do not attempt to establish whether efficient trade with respect to the total gains from trade can be achieved, but approach the problem as one of finding the trade mechanism which maximizes the gains from trade from an ex-ante perspective, finding that it matters whether one wants to maximize total or only material gains from trade, unless loss aversion is sufficiently strong.

2 Model

2.1 Utility, Social Choice Functions and Mechanisms

The set of agents is given by $I = \{S, B\}$ where S and B denote seller and buyer, respectively. It is commonly known that the type of agent $i \in I$ has distribution F_i with full support on the set $\Theta_i = [a_i, b_i] \subset \mathbb{R}_+$, and is private information. Let $\Theta = \Theta_S \times \Theta_B$ and assume that Θ_S and Θ_B have a non-trivial intersection. We interpret the type of an agent as her valuation of the good.[9] A social alternative is given by $\mathbf{x} = (y, t_S, t_B) \in X = \{0, 1\} \times \mathbb{R}^2$, where y indicates whether or not trade takes place and t_S and t_B denote the respective transfers of the seller and buyer.

Following KR, we allow for the agents to be loss averse in the trade and in the money dimension. That is, the buyer derives the standard material utility from obtaining and paying for the good, and additionally, the buyer feels weighted gain-loss utility with respect to getting the good as well as weighted gain-loss utility with respect to paying for the good. Loss-aversion is captured by value functions in the sense of Kahneman and Tversky (1979) given by

$$\mu_i^k(x) = \begin{cases} x & \text{if } x \geq 0, \\ \lambda_i^k x & \text{else,} \end{cases}$$

for some $\lambda_i^k > 1$, which reflects the degree of loss aversion.[10] Thus, the riskless total utility

not depend on other agents' reports, while we only obtain interim deterministic transfers with narrow bracketing. The setting in Gershkov et al. (2021) is more closely related to the part in Eisenhuth (2019) with wide bracketing.

[9] We could alternatively assume that the seller does not own the good but has to produce it. The seller's type would then represent her marginal cost of production. All the results that follow would go through in this case.

[10] We follow the literature by abstracting from diminishing sensitivity. This assumption is not needed for gain-loss utility in the money dimension. For instance, all the proofs go through directly if we assume $\mu_i^2(x) = g(x)$ if $x \geq 0$, and $\mu_i^2(x) = -\lambda_i^2 g(-x)$ if $x < 0$, for some concave function g. In the trade dimension, however, we cannot dispense of the piece-wise linearity, as this ensures that expected utility remains linear in the agents type in the presence of loss aversion.

is given by

$$u_S(\mathbf{x}, \mathbf{r}_S, \theta_S) = (1-y)\theta_S + t_S + \eta_S^1\mu_S^1\left(r_S^1\theta_S - y\theta_S\right) + \eta_S^2\mu_S^2(t_S - r_S^2) \tag{1}$$

$$u_B(\mathbf{x}, \mathbf{r}_B, \theta_B) = y\theta_B - t_B + \eta_B^1\mu_B^1(y\theta_B - r_B^1\theta_B) + \eta_B^2\mu_B^2(r_B^2 - t_B) \tag{2}$$

where $\eta_i^k \geq 0$ are the weights put on gain-loss utility and $\mathbf{r}_i = \{r_i^1, r_i^2\} \in \mathbb{R}^2$ are the so-called riskless reference levels. Following KR we will allow the reference point to be the agent's rational expectations and therefore a probability distribution over all riskless reference levels (see more below).

The model by KR has several moving parts, so we devote the following paragraph to discuss several (implicit) assumptions. We refer to $(1-y)\theta_S + t_S$ and $y\theta_B - t_B$ as material utility and to the other terms as gain-loss utility in the trade and money dimension, respectively. We follow most of the literature working with the model by KR and adopt the following assumption by Herweg, Müller, and Weinschenk (2010).[11]

Assumption 1 (No Dominance of Gain-Loss Utility) $\Lambda_i = \eta_i^1(\lambda_i^1 - 1) \leq 1,\ i \in I.$

As KR noted, this condition ensures that agents will not choose stochastically dominated options and the condition seems to hold up in empirical estimates.[12] Essentially, we need the assumption in order to ensure incentive compatibility and will discuss its role when stating our results.[13] Further, we follow KR by assuming "narrow bracketing", i.e., we assume that there is a separate gain-loss term for each of the two material utility dimensions, trade and money utility. This assumption is well-supported empirically (see e.g., Thaler, 1999) and is important in our setting, as it allows us to maintain quasi linearity in the presence of loss aversion.[14] Finally, the assumption that the loss aversion parameters are commonly known may seem restrictive. However, we are essentially assuming that the functional form of the utility function is common knowledge and that all private information pertains to the agents' valuation of the good. We are thereby following for instance Maskin and Riley (1984) who assume in their study of optimal auctions with risk-averse buyers that the buyers' parameter of risk-aversion is commonly known. We briefly discuss relaxing the assumption in the conclusion.

[11] This condition is commonly imposed, see for instance de Meza and Webb (2007), Eisenhuth and Grunewald (2018), Eisenhuth (2019), Karle and Peitz (2014), Rosato (2021), and Gershkov et al. (2021). For examples not adopting the assumption see Meisner and von Wangenheim (2021) or Dreyfuss, Heffetz, and Rabin (2019).

[12] In a recent meta analysis, Brown, Imai, Vieider, and Camerer (2021) find that the the loss-aversion coefficient λ is empirically estimated with a mean $\lambda = 1.955$ and a 95%-credible interval of [1.824, 2.104], suggesting that the assumption is indeed widely, if not always, satisfied.

[13] Note that the assumption applies only to gain-loss utility in the trade dimension, while no restrictions are placed on the money dimension.

[14] As discussed in the literature review above (see footnote 8 in particular), this is a key distinction to the models considering risk aversion, which implicitly correspond to wide bracketing.

A social choice function (SCF) $f : \Theta \to X$ assigns a collective choice $f(\theta_S, \theta_B) \in X$ to each possible profile of the agents' types $(\theta_S, \theta_B) \in \Theta$. In the present bilateral trade setting, a social choice function takes the form $f = (y^f, t_S^f, t_B^f)$. Let $\mathcal{F}$ denote the set of all SCFs and $\mathcal{Y}$ the set of all trade mechanisms, i.e., the set containing all y^f. A mechanism $\Gamma = (M_S, M_B, g)$ is a collection of message sets (M_S, M_B) and an outcome function $g : M_S \times M_B \to X$. We denote the direct mechanism by $\Gamma^d = (\Theta_S, \Theta_B, f)$. Since agents privately observe their types, they can condition their message on their type. Consequently, a pure strategy for agent i in a mechanism Γ is a function $s_i : \Theta_i \to M_i$. Note that $g(s_S(\theta_S), s_B(\theta_B)) \in X$. Let S_i denote the set of all pure strategies of agent i. Further, we denote the truthful strategy $s_i^t(\theta_i) = \theta_i$. Throughout, the operator $\mathbb{E}_{-i}$ denotes the expectation over the random variables $\tilde{\theta}_{-i}$ taking the value θ_i as given.

2.2 Equilibrium Concept and Revelation Principle

We use the concept of an (interim) choice-acclimating personal equilibrium (CPE) introduced in Kőszegi and Rabin (2007).[15] The set of all riskless reference levels is given by the set of all social alternatives X. Essentially, the set X captures all the outcomes that could materialize at the end of the agents' interaction. In a mechanism Γ, agent i's action induces a distribution over the set of social alternatives X, conditional on the other agent playing s_{-i}. It is this endogenously generated distribution over X that forms the agent's reference point, or rather, reference distribution in a CPE. Effectively, when an agent evaluates an outcome, she is comparing it to all other possible social alternatives that could have materialized given the distribution induced over them. Moreover, when the agent takes an action in a CPE, she takes the action anticipating that it will not only determine the outcome of the mechanism, but also the distribution over the set X and, therefore, the reference point.

Moving to the interim stage and allowing the reference point to be the agent's rational expectations, we can define the interim expected utility of the seller with type θ_S, in the mechanism Γ, when playing action $m \in M_B$, given that the buyer plays strategy s_B as

$$
\begin{aligned}
U_S(m, s_B, \Gamma | \theta_S) = & \\
& \int_{a_B}^{b_B} (1 - y^g(m, s_B(\theta_B)))\theta_S + t_S^g(m, s_B(\theta_B))\, dF_B(\theta_B) \\
& + \int_{a_B}^{b_B} \int_{a_B}^{b_B} \eta_S^1 \mu_S^1 \left(y^g(m, s_B(\theta_B'))\theta_S - y^g(m, s_B(\theta_B))\theta_S \right)\, dF_B(\theta_B')\, dF_B(\theta_B) \qquad (3)
\end{aligned}
$$

[15] KR also introduce the unacclimating personal equilibrium (UPE). In the UPE the agent "maximizes expected utility taking the reference point as given", whereas in the CPE the agent "maximizes expected utility given that it determines both the reference lottery and the outcome lottery". KR note that the CPE is more appropriate when the uncertainty is resolved after the agent's decision. We thus believe that the CPE is the more natural equilibrium concept in our context, as the report of an agent determines the uncertainty she feels about the outcome given her beliefs about the other agent's type.

$$
\begin{aligned}
&+ \int_{a_B}^{b_B} \int_{a_B}^{b_B} \eta_S^2 \mu_S^2 \left(t^g(m, s_B(\theta_B)) - t^g(m, s_B(\theta'_B)) \right) \, dF_B(\theta'_B) \, dF_B(\theta_B) \\
&= \theta_S \int_{a_B}^{b_B} (1 - y^g(m, s_B(\theta_B))) \, dF_B(\theta_B) + \int_{a_B}^{b_B} t_S^g(m, s_B(\theta_B)) \, dF_B(\theta_B) \\
&+ \theta_S \eta_S^1 \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_S^1 \left(y^g(m, s_B(\theta'_B)) - y^g(m, s_B(\theta_B)) \right) \, dF_B(\theta'_B) \, dF_B(\theta_B) \\
&+ \eta_S^2 \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_S^2 \left(t_S^g(m, s_B(\theta_B)) - t_S^g(m, s_B(\theta'_B)) \right) \, dF_B(\theta'_B) \, dF_B(\theta_B).
\end{aligned}
$$

The expression in (3) may require some explanation. The first line corresponds to material utility, the second to gain-loss utility in the trade dimension and the third to gain-loss utility in the money dimension. The double integral has a clear intuition. To illustrate, consider the last line containing the money gain-loss utility. Fix any θ_B in the domain of integration of the outer integral and suppose this was the actual realization of the buyer's type. The seller would then receive a transfer of $t_S^g(m, s_B(\theta_B))$, which she would compare to the reference point. The reference point is induced endogenously and corresponds to the distribution of possible transfers. Thus, for every θ'_B in the domain of the inner integral we get a possible transfer $t_S^g(m, s_B(\theta'_B))$ given the buyer's strategy and the seller's message. The seller compares the actual transfer $t_S^g(m, s_B(\theta_B))$ with all these other possible transfers and the value function μ_S^2 weights these comparisons differently, depending on whether they result in a loss or a gain. The inner integral then aggregates the gains and loss weighted by the induced probability distribution. Next, integrate over all the values θ_B in the domain of the outer integral to get the familiar interim expected utility. In summary, the seller aggregates over each possible realization of transfers and for each of these possible realizations she compares the outcome with all other possible outcomes, aggregating gains and losses in each comparison.

Given our interpretation that the seller owns the good, her outside option is type-dependent and given by θ_S. To simplify notation later, we will consider the seller's net utility from trade, which, with some abuse of notation, allows us to compactly write $U_S(m, s_B, \Gamma|\theta_S) = -\theta_S \tilde{v}_S(m) + \tilde{t}_S(m)$, where

$$
\begin{aligned}
\tilde{v}_S(m) &= \int_{a_B}^{b_B} y^g(m, s_B(\theta_B)) \, dF_B(\theta_B) \\
&\quad - \eta_S^1 \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_S^1 \left(y^g(m, s_B(\theta'_B)) - y^g(m, s_B(\theta_B)) \right) \, dF_B(\theta'_B) \, dF_B(\theta_B), \\
\tilde{t}_S(m) &= \int_{a_B}^{b_B} t_S^g(m, s_B(\theta_B)) \, dF_B(\theta_B) \\
&\quad + \eta_S^2 \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_S^2 \left(t_S^g(m, s_B(\theta_B)) - t_S^g(m, s_B(\theta'_B)) \right) \, dF_B(\theta'_B) \, dF_B(\theta_B).
\end{aligned}
$$

This compact notation highlights the fact that not only material utility, but also overall utility is linear in the type. Moreover, it will turn out to be useful to further define

$$\bar{t}_S(m) = \int_{a_B}^{b_B} t_S^g(m, s_B(\theta_B))\, dF_B(\theta_B),$$
$$w_S(m) = \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_B^2 \left(t_S^g(m, s_B(\theta_B)) - t_S^g(s_S(\theta_S), s_B(\theta'_B))\right)\, dF_B(\theta'_B)\, dF_B(\theta_B),$$

allowing us to write $\tilde{t}_S(m) = \bar{t}_S(m) + \eta_S^2 w_S(m)$. Similarly, we can write the buyer's utility as $U_B(m, s_S, \Gamma|\theta_B) = \theta_B \tilde{v}_B(m) + \tilde{t}_B(m)$, defining the functions $\tilde{v}_B$ and $\tilde{t}_B$ analogously.

We can now define our equilibrium concept, which follows Eisenhuth (2019).[16]

Definition 1 *A strategy profile $s^* = (s_S^*, s_B^*)$ is a CPE of the mechanism $\Gamma = (M_S, M_B, g)$ if $s_i^*(\theta_i) \in \arg\max_{m_i \in M_i} U_i(m_i, s_{-i}^*, \Gamma|\theta_i)$ for all $i \in I$ and $\theta_i \in \Theta_i$.*

Definition 2 *A mechanism Γ implements a SCF f if there is a CPE strategy profile $s = (s_S, s_B)$ such that $g(s_S(\theta_S), s_B(\theta_B)) = f(\theta_S, \theta_B)$ for all $(\theta_S, \theta_B) \in \Theta$.*

Definition 3 *A SCF f is CPE incentive compatible (CPEIC) if the truthful profile $s^t = (s_S^t, s_B^t)$ is a CPE strategy in the direct mechanism Γ^d.*

As a first result we note that the revelation principle for CPE holds in our setting.[17]

Proposition 1 (Revelation Principle for CPE) *A social choice function f can be implemented in CPE by some mechanism Γ if and only if f is CPEIC.*

The standard proof of the revelation principle goes through in spite of the presence of an endogenous reference point. To see this, note that the reference point is determined as the rational expectations over outcomes. Starting from an arbitrary mechanism which induces some distribution of outcomes, the corresponding direct mechanism induces the same distribution of outcomes and therefore also the same reference point. Henceforth, we focus on direct mechanisms and no longer explicitly list the mechanism as an argument in the utility function.

[16] In later work than Eisenhuth (2019), Dato, Grunewald, Müller, and Strack (2017) have developed a framework to extend the equilibrium concepts in Kőszegi and Rabin (2006, 2007) to study strategic interaction in finite games. The equilibrium concept they define for the CPE coincides with the one in Eisenhuth (2019) and here. Interestingly, they show that in a CPE players are unwilling to randomize over pure strategies, implying that existence may fail and that restriction to pure strategies is without loss.

[17] Proofs are relegated to the appendix unless noted otherwise.

2.3 Incentive Compatibility and Efficiency

In this section we characterize the set of all CPEIC social choice functions and introduce some familiar concepts, such as individual rationality and ex post budget balance. Further, we introduce our notion of an interim deterministic mechanism.[18]

Proposition 2 *The SCF $f = (y^f, t_S^f, t_B^f)$ is CPEIC if and only if,*

(i) $\tilde{v}_S$ is non-increasing and $\tilde{v}_B$ is non-decreasing, and

(ii) we can write utility as

$$U_S(\theta_S, s_B^t|\theta_S) = U_S(b_S, s_B^t|b_S) + \int_{\theta_S}^{b_S} \tilde{v}_S(t)\, dt, \tag{4}$$

$$U_B(\theta_B, s_S^t|\theta_B) = U_B(a_B, s_S^t|a_B) + \int_{a_B}^{\theta_B} \tilde{v}_B(t)\, dt. \tag{5}$$

Recall that the functions $\tilde{v}_B$ and $\tilde{v}_S$ contain terms of gain-loss utility. Thus, while the incentive-compatibility conditions in Proposition 2 *seem* similar as those in the absence of loss aversion, they are not and thus the set of incentive-compatible SCF need not coincide either. We say that a SCF is individually rational if for both agents $i \in I$

$$U_i(\theta_i, s_{-i}^t|\theta_i) \geq 0 \quad \forall \theta_i \in \Theta_i. \tag{IR}$$

Setting the outside option in (IR) equal to zero is without loss of generality.[19] An agent could choose to walk away and not participate in the mechanism as soon as she learns her type. Doing so would rule out any possibility of trade and payment or receipt of any transfers. Therefore, the reference points of the agent would be equal to zero, as she anticipates that no trade or transfers can take place if she walks away. Consequently, there would be no feelings of gain or loss, as well as zero material utility.

We say that a mechanism has interim-deterministic transfers, when, given her own type, an agent's transfer does not depend on almost all types of the other agent. Similarly, a trade rule is interim deterministic, when, given her own type, the trade rule coincides for almost all types of the other agent. A mechanism with interim-deterministic transfers and an interim-deterministic trade rule is called interim deterministic.

[18] In contrast to Carbajal and Ely (2016), who consider price discrimination using a different model of loss aversion than the one here, the standard integral representation obtains in our setting. This is driven by the fact that, in contrast to Carbajal and Ely (2016), the report of an agent and not her type determines her reference point. For instance, a high buyer type does not expect to get the good with the probability corresponding to her true type when misreporting. Rather, she is aware that reporting a lower type changes the probability of getting the good and this is reflected in her reference point.

[19] Recall that we are considering net utility and have thus already taken care of the seller's type-dependent outside option.

3 Attachment, Endowment and Information Rents

As noted in the introduction, the attachment and endowment effect have been empirically documented in bilateral trade situations. However, the classical model with quasi-linear utility as in Myerson and Satterthwaite (1983) cannot explain such effects, motivating the inclusion of expectations-based loss aversion in the present paper. Our first step is thus to formally identify these effects and their implications in our model.

Proposition 3 *In any CPEIC mechanism, the information rent of the seller is increasing in Λ_S and the information rent of the buyer is decreasing in Λ_B.*

Put differently, the presence of loss aversion in the trade dimension increases the information rent of the seller and decreases the information rent of the buyer. The proof is straightforward as it suffices to take the derivatives with respect to Λ_S and Λ_B from equations (4) and (5), respectively. The key step is to note that

$$\begin{aligned}
&\tilde{v}_B(\theta_B) \\
&= \int_{a_S}^{b_S} y^f(\theta_S,\theta_B)\, dF_S(\theta_S) + \eta_B^1 \int_{a_S}^{b_S}\int_{a_S}^{b_S} \mu_B^1 \left(y^f(\theta_S,\theta_B) - y^f(\theta'_S,\theta_B)\right)\, dF_S(\theta'_S)\, dF_S(\theta_S), \\
&= y_B(\theta_B) + \eta_B^1 \int_{a_S}^{b_S}\int_{a_S}^{b_S} y^f(\theta_S,\theta_B)(1-y^f(\theta'_S,\theta_B)) - \lambda_B^1(1-y^f(\theta_S,\theta_B))y^f(\theta'_S,\theta_B)\, dF_S(\theta'_S)\, dF_S(\theta_S), \\
&= y_B(\theta_B)(1-\Lambda_B(1-y_B(\theta_B)))
\end{aligned}$$

and analogously for the seller $\tilde{v}_S(\theta_S) = y_S(\theta_S)(1+\Lambda_S(1-y_S(\theta_S)))$, where

$$y_B(\theta_B) = \int_{a_S}^{b_S} y^f(\theta_S,\theta_B) dF_S(\theta_S), \quad y_S(\theta_S) = \int_{a_B}^{b_B} y^f(\theta_S,\theta_B) dF_B(\theta_B).$$

Thus, the attachment effect is captured by $-\Lambda_B \int_{a_B}^{\theta_B} y_B(\theta_B)(1-y_B(\theta_B)d\theta_B$ and the attachment effect by $\Lambda_S \int_{\theta_S}^{b_S} y_S(\theta_S)(1-y_S(\theta_S))d\theta_S$. As already noted, the respective decrease and increase in the rents stem from the loss aversion in the trade dimension. As one would expect, loss aversion on the money dimension plays no role here. To simplify exposition, we will use the term loss aversion as referring to loss aversion in the trade dimension unless stated differently.

Having formally identified the two effects as the impact of loss aversion on the information rents, we can conduct an interesting bit of comparative statics. Does loss aversion affect all types in the same way?

Corollary 1 *The strength of the endowment and attachment effect is increasing and decreasing in the type of the buyer and seller, respectively.*

The result follows immediately as one takes the derivative with respect to Λ_S and θ_S from equation (4) and with respect to Λ_B and θ_B from equation (5), so that a proof

is omitted. The finding about the impact of the endowment and attachment effect on the information rents suggests that the presence of a loss-averse buyer could enable the designer to implement materially efficient trade subject to ex-post budget balance and the agents' participation constraints,[20] that is, to "reverse" the impossibility result by Myerson and Satterthwaite (1983). To see this, recall the interpretation of the result, stating that the gains from trade do not suffice to cover the agents' information rents. Thus, seller loss aversion, which increases the information rent, will make the problem only harder, while buyer loss aversion could make it easier. However, there is a countervailing effect even for the buyer. Loss aversion not only affects information rents as stated in Proposition 3, but also decreases expected utility and hence makes satisfying the participation constraints harder, too.[21] Nevertheless, it suffices to consider buyer loss aversion to check whether the impossibility result can be reversed. Making use of this insight, we can proceed analogously to the proof in Myerson and Satterthwaite (1983). That is, impose budget balance as well as incentive compatibility to obtain an expression for the sum of utilities of the "worst" buyer and seller types in the materially efficient mechanism and show that it is strictly negative. Indeed, we obtain

$$\begin{aligned} & U_B(a_B) + U_S(b_S) = \\ & - \int_{\max\{a_B, a_S\}}^{\min\{b_S, b_B\}} (1 - F_B(x))F_S(x)(1 - \Lambda_B(1 - F_S(x))) + \Lambda_B(1 - F_S(x))F_S(x)x f_B(x)\, dx \qquad (6) \\ & < 0, \end{aligned}$$

which violates individual rationality for any $\Lambda_B \leq 1$. This proves our next result (see Appendix A for the details).

Proposition 4 *Given CPEIC, individual rationality and ex-post budget balance, it is impossible to realize all material gains from trade for any degree of loss aversion in the money or trade dimension.*

The minimal subsidy needed to induce materially efficient trade under CPEIC and IR in equation (6) can be interpreted as a measure of the severity of the impossibility problem and will generally depend on the degree of loss aversion and the distribution of the agents' types. Indeed, taking the derivative of the minimal subsidy in equation (6) with respect to Λ_B, we can see that the attachment effect mitigates the impossibility problem by dominating the diminishing effect of loss aversion on the participation constraints

[20] Ex-post budget balance corresponds to the condition $t_S^f(\theta_S, \theta_B) = t_B^f(\theta_S, \theta_B), \quad \forall(\theta_S, \theta_B) \in \Theta$.

[21] Loss aversion on the money dimension does not affect information rents but also reduces expected utility, thus only making it harder to reverse the impossibility result.

whenever

$$\int_{\max\{a_B,a_S\}}^{\min\{b_S,b_B\}} (1-F_B(x))F_S(x)(1-F_S(x)) - (1-F_S(x))F_S(x)xf_B(x)\,dx \geq 0.$$

To get a feel for this condition, consider the families of distributions $F_S(x) = x^s$ and $F_B(x) = x^b$ on $[0,1]$ for $b, s > 0$. Whenever $b > 2s^2 - 1$ the buyer's loss aversion makes the problem easier. In words, the likelier low seller types and high buyer types are, the less severe is the impossibility problem. This is in line with the intuition underlying the attachment effect. When low seller types are likely, a buyer puts a relatively high probability on trade taking place and thus has a strong attachment to the good (a high reference point). Hence, when low seller types and high buyer types are likely, on average the buyer will have a high attachment effect, thereby mitigating the impossibility problem. Note that in the absence of loss aversion, it is also true that the minimal subsidy is lower the likelier low seller types and high buyer types are. In the presence of the attachment effect, however, this is reinforced.

Another noteworthy point is that for the extreme types, i.e., types who lie outside the intersection of the intervals, loss aversion does not matter. This finding is very intuitive. To see this, observe that for these types trade is interim deterministic and hence there is no gain-loss utility as there is no room for ex-post variations in payoffs. Put differently, expectations-based loss aversion only has bite when there is unresolved uncertainty, which is only the case for types lying strictly in the intersection of the type spaces.

The fact that the impossibility result is not reversed is linked to the assumption that $\Lambda_B \leq 1$, i.e., that gain-loss utility does not dominate for the buyer. For instance, when types are drawn from $[0,1]$ with distributions $F_S(x) = x$ and $F_B(x) = x^{10}$ the subsidy in equation (6) turns into a surplus for $\Lambda_B \geq 13/3$. However, in this example $\Lambda_B \leq 1$ is a necessary condition for the materially efficient mechanism to be incentive compatible for the buyer. Hence, incentive compatibility puts limits on the feasible degree of loss aversion, and, as a consequence, on the strength of the attachment effect, meaning that the impossibility result cannot be reversed. Yet, as we will discuss next, $\Lambda_B \leq 1$ is in general only a sufficient condition for incentive compatibility and not always necessary.

The assumption that $\Lambda_i \leq 1$ is commonly imposed in the literature for conceptual as well as technical reasons and seems appears widely supported empirically (see footnotes 11 and 12). In particular, KR showed that the assumption ensures that agents do not choose stochastically dominated options. In the present context, it is easy to show that the assumption is a sufficient condition for the materially efficient trade rule to be incentive compatible in the presence of loss aversion. Moreover, whenever $F_S(a_B) = 0$ the assumption is not only sufficient, but also necessary. That is, whenever the smallest buyer type has a zero probability of trading, the materially efficient trading rule is

CPEIC if and only if $\Lambda_B \leq 1$. In particular, this is true when the types of both agents are drawn from the same support. It turns out, however, that when $F_S(a_B) > 0$ the assumption is no longer necessary.[22] Indeed, when $F_S(a_B) < 1/2$ the necessary condition reads $\Lambda_B \leq 1/(1 - 2F_S(a_B))$ and when $F_S(a_B) \geq 1/2$ no restrictions need to be put on Λ_B. In the light of the above result the question thus arises whether the impossibility result persists when $F_S(a_B) > 0$ and the assumption is relaxed, as this would allow us to strengthen the attachment effect and possibly set the required subsidy in equation (6) equal to zero.

To this end, one can show that the impossibility result continues to hold for $\Lambda_B \leq 1/(1-F_S(a_B))$. This condition ensures that the lowest buyer type a_B is in fact the "worst" buyer type. For $\Lambda_B > 1/(1 - F_S(a_B))$, the worst buyer type is some intermediate type and the above approach to proving the impossibility result fails: if the lowest buyer type is no longer the worst type, satisfying individual rationality for the lowest buyer type does no longer guarantee satisfying individual rationality for all types. The observation that an intermediate type is the worst type is reminiscent of the related model of partnership dissolution (Cramton, Gibbons, and Klemperer, 1987; Fieseler, Kittsteiner, and Moldovanu, 2003). In this model, the good is initially not exclusively owned by one agent only, but by several agents. As a result, the worst type of an agent may be an intermediate type. However, in spite of this similarity, the approach taken in that model cannot be extended to the present context due to the endogeneity of the reference point. In sum, although counterexamples have proved elusive, a reversal of the impossibility for when $\Lambda_B > 1/(1-F_S(a_B))$ cannot be ruled out. Note, however, that for sufficiently high degrees of loss aversion the total gains from trade disappear completely. Thus, even if the buyer's information rent can be reduced using the attachment effect, impossibility will obtain for sufficiently high degrees of loss aversion because it will eliminate all the total gains from trade.[23]

4 Optimal Mechanisms

The preceding section has formally identified the endowment and the attachment effect in an otherwise standard bilateral trade setting. In particular, we have seen how loss aversion impacts the agents' information rents and the participation constraints, allowing us to show that the impossibility of implementing materially efficient trade extends to

[22] In Herweg et al. (2010), who first introduced this assumption, the assumption plays a similar role as here. It provides a sufficient but not necessary condition to satisfy incentive compatibility of certain contracts.

[23] In the above we have only discussed the degree of loss aversion of the buyer. Analogous arguments regarding the necessity and sufficiency of $\Lambda_S \leq 1$ for incentive compatibility of the seller apply. However, as loss aversion on the side of the seller makes the impossibility problem only harder, it does not enter our result.

the setting with loss-averse agents. We now turn to the problem of designing optimal mechanisms. We begin by considering the problem of maximizing the designer's revenue and then turn to the (conceptually) more nuanced question of maximizing the gains from trade. In contrast to the previous section, we assume a symmetric support for the distributions of buyer and seller types to simplify notation.[24] We will continue to allow for arbitrary distributions, but will make the standard regularity assumption of increasing virtual types.

4.1 Maximizing the Designer's Revenue

The revenue-maximizing designer's problem reads

$$\max_{(y^f,t_S^f,t_B^f)\in\mathcal{F}} \int_a^b \int_a^b \left(t_B^f(\theta_S,\theta_B) - t_S^f(\theta_S,\theta_B)\right)\, dF_S(\theta_S)\, dF_B(\theta_B),$$

subject to CPEIC and IR. (RM)

We begin by rewriting this problem into a more accessible form which will allow us to gain some intuition first. The first step is to impose the envelope representation of the utility due to the CPEIC and the individual rationality constraint. The objective function then reads

$$\begin{aligned}&\int_a^b \left(\eta_B^2 w_B(\theta_B) + \theta_B \tilde{v}_B(\theta_B) - \int_a^{\theta_B} \tilde{v}_B(t)\, dt\right)\, dF_B(\theta_B)\\ &+ \int_a^b \left(\eta_S^2 w_S(\theta_S) - \theta_S \tilde{v}_S(\theta_S) - \int_{\theta_S}^b \tilde{v}_S(t)\, dt\right)\, dF_S(\theta_S). \qquad (7)\end{aligned}$$

In the absence of loss aversion, the envelope representation of utility would allow us to maximize over the trade rule only instead of both the trade rule and transfers. With loss aversion in the money dimension, however, this is not the case. Indeed, recall that we defined

$$w_S(\theta_S) = \int_a^b \int_a^b \mu_S^2 \left(t_S^f(\theta_S,\theta_B) - t_S^f(\theta_S,\theta_B')\right)\, dF_B(\theta_B')\, dF_B(\theta_B),$$

and thus the objective function still depends on transfers. This expression and its analog for the buyer collect all gain-loss utility with respect to money. Nevertheless, the problem can be reduced to only choosing the optimal trade rule, because in any optimal mechanism the transfers of the seller will be interim deterministic and thus not depend on the buyer's type, and vice versa, so that $w_i(\theta_i) = 0$.

[24] All arguments go through analogously for the case with asymmetric supports.

Proposition 5 *Any solution to the revenue maximization problem* (RM) *entails interim-deterministic transfers.*

Intuitively, loss-averse agents dislike ex-post variations in their payoffs. By making the transfers independent of the other agent's type, the designer completely insures the agents from any ex-post variation in the transfers. Thus, starting from any mechanism with non-interim-deterministic transfers, the designer can extract more surplus from the agents by choosing appropriate interim-deterministic transfers, effectively selling the agents insurance. Note that interim-deterministic transfers are also *a* solution in the absence of loss aversion. However, in the presence of loss aversion interim-deterministic transfers are the *only* solution.[25]

Proposition 5 allows us to rewrite the maximization problem to

$$\max_{y^f \in \mathcal{Y}} \int_a^b J_B(\theta_B) y_B(\theta_B)(1 - \Lambda_B\,(1 - y_B(\theta_B))\, f_B(\theta_B) d\theta_B \\ - \int_a^b J_S(\theta_S) y_S(\theta_S)(1 + \Lambda_S(1 - y_S(\theta_S)))f_S(\theta_S) d\theta_S \qquad \text{(RM')}$$

subject to $y_B(\theta_B)$ being non-decreasing and $y_S(\theta_S)$ being non-increasing,

where $y_B(\theta_B) = \int_a^b y^f(\theta_S, \theta_B)\ dF_S(\theta_S)$ and $y_S(\theta_S) = \int_a^b y^f(\theta_S, \theta_B)\ dF_B(\theta_B)$ denote the interim trade probabilities of the buyer and seller, respectively, and

$$J_B(\theta_B) = \theta_B - \frac{1 - F_B(\theta_B)}{f_B(\theta_B)}, \quad J_S(\theta_S) = \theta_S + \frac{F_S(\theta_S)}{f_S(\theta_S)}$$

denote the buyer's and the seller's virtual types. We make the following standard assumption.

Assumption 2 (Regularity) *The virtual types J_B and J_S are strictly increasing.*

The designer faces the trade-off that inducing trade comes at a cost in the form of the payment due to the seller and with a benefit in the form of the payment from the buyer. Further, the form of the objective function in (RM') suggests that even in the presence of loss aversion the designer wants to induce trade between high buyer and low seller types in particular. Put differently, the designer wants to buy the good from a low-value seller and sell it to a high-value buyer, as this yields a large profit margin. However, as a consequence

[25] Eisenhuth (2019) proved an analogous result for the case of auctions. In fact, one can show that Proposition 5 extends beyond the bilateral trade and auction setting. Further, the result is reminiscent of the optimal mechanism found in Herweg et al. (2010), who augment a principal-agent setting with moral hazard by assuming the agent is expectations-based loss averse as in the present paper. They find that the principal optimally employs a binary payment scheme instead of a fully contingent contract in the presence of loss aversion. Hence, loss aversion drastically reduces the ex-post variation in payments, too, but, in contrast to the present setting, does not eliminate it fully to preserve incentives.

of expectations-based loss aversion, it matters for an agent's utility whether trade takes place with only a few or many types of the other agent, as this affects her expectations, which in turn affect the strength of the endowment and attachment effect. Thus, there are in some sense externalities between the outcomes of different types. Indeed, because the agents' expected utilities endogenously depend on the mechanism through the reference point, point-wise maximization of the objective function is not possible. To see this, note that the expected trade probabilities y_B and y_S enter both linearly and quadratically so that we cannot "move out" the integral of y_B and y_B to maximize over the ex-post trade rule y^f. In order to get rid of the quadratic terms, we first show that the optimal trade rule takes a particular form.

We begin by performing a change of variables, which will simplify the analysis. Let $v_i = F_i(\theta_i)$ and define $\varphi_i(v_i) = F_i^{-1}(v_i)$. Further, define

$$q(v_B, v_S) = y(\varphi_B(v_B), \varphi_S(v_S)), \quad q_i(v_i) = \int_0^1 q(v_B, v_S) dt_{-i}$$

as well as

$$\begin{aligned} M_B(v_B) &= J_B(\varphi_B(v_B)) = \varphi_B(v_B) - (1 - v_B)\varphi_B'(v_B), \\ M_S(v_S) &= J_S(\varphi_S(v_S)) = \varphi_S(v_S) + v_S\varphi_S'(v_S). \end{aligned}$$

The problem then becomes

$$\begin{aligned} &\int_0^1 M_B(v_B) q_B(v_B)(1 - \Lambda_B(1 - q_B(v_B)))dv_B \\ &- \int_0^1 M_S(v_S) q_S(v_S)(1 + \Lambda_S(1 - q_S(v_S))dv_S \end{aligned} \tag{8}$$

subject to the monotonicity constraints and we obtain the following intermediate result.

Lemma 1 *The solution to the problem* (8) *can be written as*

$$q(v_B, v_S) = \begin{cases} 1 & 0 \leq v_S \leq v_S^*(v_B) \\ 0 & o.w. \end{cases} \tag{9}$$

for some function $v_S^* : [0,1] \to [0,1]$.

The above lemma has a straight-forward interpretation. Fix a buyer type v_B and suppose that it is optimal to induce trade for some seller type $v_S^*(v_B)$. Then, it is optimal to also induce for all seller types v_S that are lower, i.e., for all $v_S \leq v_S^*(v_B)$. This reflects the intuition that the designer would like to induce trade with low seller types, as they will be willing to give up the good at a low price.

The proof of the lemma is in the appendix and proceeds in three steps. First, suppose some trade rule $\hat{q}$ is optimal and associate to $\hat{q}$ the function

$$q(v_B, v_S) = \begin{cases} 1 & 0 \leq v_S \leq \hat{q}_B(v_B) \\ 0 & o.w., \end{cases} \tag{10}$$

where $\hat{q}_B(v_B) = \int_0^1 \hat{q}(v_B, v_S) dv_S$. Note that $q_B = \hat{q}_B$ by construction so that the first integral equation (8) is not affected by a change from $\hat{q}$ to q. Essentially, we are holding fixed the expected trade probability of the buyer and shift all trade probability to low seller types. Second, we prove a technical lemma in the appendix (Lemma 2) allowing us to show that

$$\int_0^1 M_S(v_S) q_S(v_S)(1 + \Lambda_S) dv_S \leq \int_0^1 M_S(v_S) \hat{q}_S(v_S)(1 + \Lambda_S) dv_S \tag{11}$$

and

$$\int_0^1 \Lambda_S M_S(v_S)(q_S^2(v_S) - \hat{q}_S^2(v_S)) dv_S \geq 0. \tag{12}$$

Third, we plug that together to show that second integral in equation (8) has become smaller, implying that q yields a higher revenue than the initial trade rule $\hat{q}$, showing that the latter cannot be optimal, completing the proof.

Making use of Lemma 1, we note that

$$q_B(v_B)^2 = \left(\int_0^1 q(v_B, v_S) dv_S \right)^2 = 2 \int_0^1 q(v_B, v_S) v_S dv_S$$

and that

$$q_S(v_S)^2 = \left(\int_0^1 q(v_B, v_S) dv_B \right)^2 = 2 \int_0^1 q(v_B, v_S)(1 - v_B) dv_B.$$

This allows us to get rid of the quadratic terms in (8) and move out the integral of the expected trade probabilities to obtain

$$\begin{aligned} &\int_0^1 \int_0^1 M_B(v_B) \left[1 - \Lambda_B + 2\Lambda_B v_S\right] q(v_B, v_S) dv_S dv_B \\ -&\int_0^1 \int_0^1 M_S(v_S) \left[1 - \Lambda_S + 2\Lambda_S v_B\right] q(v_B, v_S) dv_S dv_B, \end{aligned}$$

which, reversing our change of variables, becomes

$$\begin{aligned}
&\int_a^b \int_a^b \underbrace{J_B(\theta_B)\left[1-\Lambda_B+2\Lambda_B F_S(\theta_S)\right]}_{:=\tilde{J}_B(\theta_B,\theta_S)} y^f(\theta_B,\theta_S)dF_S(\theta_S)dF_B(\theta_B) \\
&-\int_a^b \int_a^b \underbrace{J_S(\theta_S)\left[1-\Lambda_S+2\Lambda_S F_B(\theta_B)\right]}_{\tilde{J}_S(\theta_S,\theta_B)} y^f(\theta_B,\theta_S)dF_S(\theta_S)dF_B(\theta_B), \\
&=\int_a^b \int_a^b \left(\tilde{J}_B(\theta_B,\theta_S)-\tilde{J}_S(\theta_S,\theta_B)\right) y^f(\theta_B,\theta_S)dF_S(\theta_S)dF_B(\theta_B).
\end{aligned}$$

We now have a concave maximization problem so that an trade rule y^f is optimal if and only if (see, e.g., Theorems 1 and 2 in Luenberger, 1969, p. 217 and p. 221)

$$y^f(\theta_B,\theta_S)=\begin{cases}1 & \text{if } \tilde{J}_B(\theta_B,\theta_S)-\tilde{J}_S(\theta_S,\theta_B)\geq 0\\ 0 & \text{otherwise}\end{cases} \tag{13}$$

subject to the monotonicity constraints on y_B and y_S. In order to get a more precise statement in terms of the monotonicity constraints, we make use of Assumption 1, i.e., that $\Lambda_i \leq 1$. This allows us to reformulate to

$$\begin{aligned}
&\tilde{J}_B(\theta_B,\theta_S)-\tilde{J}_S(\theta_S,\theta_B)\geq 0 \\
\Leftrightarrow &\bar{J}_B(\theta_B):=\frac{J_B(\theta_B)}{1-\Lambda_S+2\Lambda_S F_B(\theta_B)}\geq \frac{J_S(\theta_S)}{1-\Lambda_B+2\Lambda_B F_S(\theta_S)}=:\bar{J}_S(\theta_S).
\end{aligned}$$

Thus, the trade rule in equation 13 satisfies the monotonicity constraints if the functions $\bar{J}_i$ are increasing. Note that for the case $\Lambda_B=\Lambda_S=0$ we obtain $\bar{J}_i=J_i$ so that the monotonicity follows directly from the regularity assumption. In general, however, the functions $\bar{J}_i$ are not necessarily strictly increasing, so that it is not clear, weather the monotonicity constraint is satisfied. We have

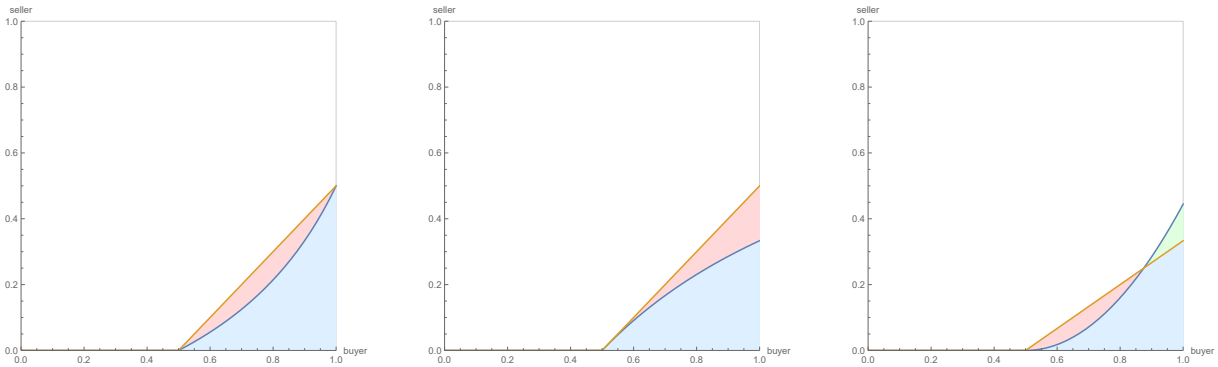
$$\frac{\partial \bar{J}_i'(\theta_i)}{\partial \theta_i}>0 \Leftrightarrow \Lambda_j < \frac{J_i'(\theta_i)}{2J_i(\theta_i)f_i(\theta_i)+J_i'(\theta_i)(1-2F_i(\theta_i))}$$

allowing us to define the set

$$IC=\left\{(\Lambda_B,\Lambda_S)\geq 0 \mid \Lambda_j < \frac{J_i'(\theta_i)}{2J_i(\theta_i)f_i(\theta_i)+J_i'(\theta_i)(1-2F_i(\theta_i))}\ \forall \theta_i\in[a,b]\right\}.$$

The above derivations prove the following result.

Proposition 6 *Suppose $(\Lambda_B,\Lambda_S)\in IC$. Then, the revenue-maximizing trade rule is given by trade taking place if and only if $\bar{J}_B(\theta_B)\geq \bar{J}_S(\theta_S,)$.*



(a) $(\Lambda_B, \Lambda_S) = (1/2, 0)$ (b) $(\Lambda_B, \Lambda_S) = (0, 1/2)$ (c) $(\Lambda_B, \Lambda_S) = (1, 0)$

Figure 1: The above figures depict optimal trade rules with the buyer and seller types on the x and y axes, respectively. Buyer types are always drawn from $F_B(\theta_B) = \theta_B$. Seller types in panels (a) and (b) are drawn from $F_S(\theta_S) = \theta_S$ and from $F_S(\theta_S) = \sqrt{\theta_S}$ in panel (b). The orange line corresponds to the case without loss aversion. The blue line corresponds to the loss aversion parameters indicated in the sub captions. For profiles (θ_B, θ_S) in the red-shaded are, loss aversion leads to a reduction of trade and in the green-shaded area to an increase in trade; in the blue-shaded area trade takes place in with and without loss aversion.

This result deserves some discussion, as it has several noteworthy features. First, in the absence of loss aversion in the trade dimension, i.e., for $\Lambda_S = \Lambda_B = 0$, we obtain the mechanism from Myerson and Satterthwaite (1983). Second, depending on the distribution of types and degrees of loss aversion, the trade frequency can increase or decrease compared to the case with no loss aversion. This is illustrated in Figure 1c, where more trade takes place for pairs of relatively high buyer and seller types, but less trade is induced for relatively low pairs. To gain some intuition, notice that low seller types are relatively likely given the chosen distributions. Thus, for high buyer types, it is fairly likely that they value the good more than the seller. Increasing the trade probability for high buyer types by also inducing trade with higher seller types therefore does not increase ex-post variation in payoffs too much. Moreover, the attachment effect makes it cheaper to induce truthful behavior among buyers, so that inducing more trade for higher types (for which the attachment effect is strongest, see Corollary 1) is attractive. In other cases, the trade frequency always decreases, as is illustrated in Figures 1a and 1b. Thus, in such cases, the designer not only provides agents with full insurance in the money dimension by means of interim-deterministic transfers, but also offers partial insurance in the trade dimension by reducing the trade probability. Finally, let us discuss the restriction we need to place on the degree of loss aversion to ensure incentive compatibility. We can look at the functions $\tilde{J}_i$ as modified virtual types. However, due to the endogeneity of the reference point, these modified virtual types also depend on the other agent's type, so that an ironing approach (Myerson, 1981) is not feasible. Thus, we need to formulate

conditions on the parameters of loss aversion that ensure CPEIC. Inspecting the set IC reveals that the above trade rule will only be optimal for sufficiently small degrees of loss aversion, often well below simply assuming no dominance of gain-loss utility, as the next part will illustrate.

To close this section, we focus our attention on the case when types are uniformly distributed on $[a, a+1]$, as this allows us to derive closed form solutions of the optimal mechanism, allowing for interesting comparative statics and sharper insights into the necessary bounds on the degree of loss aversion.

Corollary 2 *Consider the case of uniformly distributed types on the interval* $[a, a+1]$ *for* $a \geq 0$ *and suppose that* $\Lambda_B \leq 1/(a+1), \Lambda_S \leq \min\{1, 1/a\}$. *Then, the optimal trade rule reads*

$$y(\theta_S, \theta_B) = \begin{cases} 1 & \text{if } \theta_S \leq \delta(\theta_B), \\ 0 & \text{otherwise}, \end{cases}$$

where

$$\delta(\theta_B) = \frac{(2\theta_B - 1 - a)(1 - \Lambda_B(2a+1) + a\Lambda_S) + a - \Lambda_S a^2}{2(1 - \Lambda_B(2\theta_B - a - 1) + \Lambda_S(2\theta_B - 1 - 2a))}.$$

Moreover, for $(\Lambda_B, \Lambda_S) \neq (0, 0)$, *increasing the parameter* a *reduces the optimal trade frequency, eventually eliminating all trade.*

We can interpret an increase of a as an increase of the stakes. Thus, for higher stakes, less trade takes place for any positive degree of loss aversion. This is in sharp contrast to the case without loss aversion, where the optimal mechanism is independent of the size of the stakes. Intuitively, the potential material gains from trade remain the same even when the stakes are high, because only the difference between valuation matters. However, as the stakes increase, the potential losses increase. Since the designer needs to compensate the agents for these losses with appropriate transfers to maintain participation, the losses eventually eat up all the potential material gains. Hence, at some point the best the designer can do is to induce no trade at all. Contrary to conventional wisdom, the behavioral effects of loss aversion are not mitigated when the stakes are large. Rather, in that case, loss aversion has the biggest impact precisely when the stakes are large. Further, we note that the bound on loss aversion ensuring incentive compatibility decreases in the degree of loss aversion.

4.2 Maximizing the Gains from Trade

In this section, we consider the problem of maximizing gains from trade. In the absence of loss aversion, the objective function is given by the sum of ex-ante expected utilities of the two agents. In the presence of loss aversion, however, it may not be clear what constitutes an appropriate objective function. Naturally, one way to go about is to mirror the case without loss aversion and to maximize the sum of ex-ante expected utilities. But what if the designer is only interested in maximizing the material gains from trade, e.g., because she considers loss aversion a mistake?

In standard welfare economics, choice reveals a preference, which in turn should guide any welfare considerations. When choices do not reveal a preference because of inconsistencies or, mistakes, the case is not so clear. Within the field of behavioral welfare economics, we can distinguish between model-based and model-less approaches (Manzini and Mariotti, 2014). In a model-based approach the welfare criterion is developed based on an underlying theory (or, a model) of mistakes. In contrast, in a model-less approach multiple inconsistent preferences are being aggregated into a welfare criterion solely on the basis of observed choices. In analogy, when maximizing the trade from gains the designer could "take loss aversion seriously" and include gain-loss utility in the objective function, or "treat loss aversion as a mistake", thus only considering material gains from trade in the maximization problem. It is not always straightforward or uncontroversial to determine the "right" approach in such situations. As we will see, the distinction matters in general, but may be irrelevant in special cases.

In order to formulate the maximization problem, we impose a budget balance condition in addition to CPEIC and IR. Namely, we do not want the designer to inject money in the economy on average. This is in line with the preceding section, where we looked at ex-ante revenue maximization. We say that a mechanism is ex-ante budget balanced if

$$\int_a^b \int_a^b \left(t_S^f(\theta_S, \theta_B) - t_B^f(\theta_S, \theta_B) \right) \, dF_S(\theta_S) \, dF_B(\theta_B) = 0. \tag{AB}$$

We consider two maximization problems given by

$$\max_{(y^f, t_B^f, t_S^f) \in \mathcal{F}} \int_a^b U_S(\theta_S, s_B^t | \theta_S) \, dF_S(\theta_S) + \int_a^b U_B(\theta_B, s_S^t | \theta_B) \, dF_B(\theta_B),$$

$$\text{subject to CPEIC, IR and AB.} \tag{TG}$$

and

$$\max_{(y^f,t_B^f,t_S^f)\in\mathcal{F}} \int_a^b (-\theta_S y_S(\theta_S) + \bar{t}_S(\theta_S))\ dF_S(\theta_S) + \int_a^b (\theta_B y_B(\theta_B) - \bar{t}_B(\theta_B))\ dF_B(\theta_B),$$

subject to CPEIC, IR and AB. (MG)

In problem TG the designer includes gain-loss utility in the objective function and thus maximizes what we call total gains from trade, whereas only material gains from trade are maximized in problem MG. To solve either problem, we proceed as we did before and also obtain the result that in any mechanism maximizing total or material gains from trade agents are fully insured against any ex-post variation in transfers.

Proposition 7 *Any solution to the problem* (TG) *or* (MG) *entails interim-deterministic transfers.*

The proof is analogous to the revenue maximization problem and thus omitted. From here we proceed as we did for the derivation of the revenue-maximizing mechanism, the only difference being the presence of the budget constraint. Putting all of this together, we obtain the following result.[26]

Proposition 8 *Suppose* $(\Lambda_B, \Lambda_S) \in IC^j$ *for* $j \in \{TG, MG\}$. *Then, the optimal trade rules for problems (TG) and (MG), respectively, are such that trade takes place if and only if trade takes place whenever* $\bar{J}_B^j(\theta_B, \gamma) \geq \bar{J}_S^j(\theta_S, \gamma)$.

Naturally, the two trade rules coincide when $\Lambda_B = \Lambda_S = 0$, in which case we find ourselves in the setting as in MS. In general, however, the optimal trade rules (and thus transfers) for the two problems will be distinct and it will matter what stance the designer takes regarding gain-loss utility. Yet, in some instances, it does not matter whether the designer treats gain-loss utility as a mistake or not, as the following corollary shows.

Corollary 3 *Consider the case of uniformly distributed types on the unit interval. If* $\Lambda_S = \Lambda_B = \Lambda$, *then the optimal trade rules for the problems (TG) and (MG) coincide.*

5 Conclusion

The theoretical and empirical literature on bilateral trade have both become quite extensive over time. However, the theoretical literature has so far failed to incorporate some findings from the empirical literature, most prominently the well-documented endowment

[26] The parameter γ is the Lagrange multiplier. See the proof of the result for the definitions of the expressions in the result. They are analogous to the ones in the revenue-maximization problem above.

and attachment effect. The present paper aims to fill this gap by augmenting the standard model by Myerson and Satterthwaite (1983) with expectations-based loss aversion as in Kőszegi and Rabin (2006, 2007). In doing so, we also contribute to the literature combining mechanism design and loss aversion (see Kőszegi, 2014).

We first formally identify the endowment and attachment effect and study their impact on the agents' information rents. Using these insights, we can show that it remains impossible to implement ex-post materially efficient trade, but that buyer loss aversion can mitigate the severity of this impossibility. Turning to the design of optimal mechanisms we find that the designer optimally provides agents with insurance in order to reduce ex-post variation in payoffs. More specifically, when maximizing revenue or gains from trade, agents receive full insurance in the money dimension in the form of interim-deterministic transfers. In terms of the trade rule, we show that depending on the distribution of types, the trade frequency may increase or decrease in the presence of loss aversion.

One may wonder whether other models than the one by Kőszegi and Rabin can also explain the attachment and endowment effect and thus constitute alternatives to the present analysis. One obvious alternative is a model of loss aversion with a fixed reference point, such as classical prospect theory by Kahneman and Tversky. Indeed, with an appropriately chosen, fixed reference point, such a model can give rise to both attachment and endowment effect. However, the innovation of Kőszegi and Rabin was precisely to determine the reference point endogenously, as otherwise the question of what the appropriate reference point is, remains open. Yet, even with an endogenously determined reference point there exist alternative ways to proceed. Kőszegi and Rabin (2007) note that the models of disappointment aversion by Bell (1985) and Loomes and Sugden (1986) are very similar except that the endogenous reference point is given by the certainty equivalent of a lottery rather than the full lottery. However, Masatlioglu and Raymond (2016) find that the intersection of preferences induced by expectations-based loss aversion with CPE and any of these disappointment-aversion models is only standard expected utility, and thus while seemingly similar, the models are actually quite different. Nevertheless, Benkert (2022) shows that the optimal mechanisms for the two types of models are equivalent across a range of mechanism design settings. In particular, the optimal mechanisms derived in the present paper are also optimal if we instead work with a model of disappointment aversion as in Bell (1985) and Loomes and Sugden (1986). This finding is of practical relevance, as the designer of some economic institution may have evidence that individuals are loss averse, but be unsure about the precise formation process of the reference point, be it fixed, as a full lottery over outcomes or as the certainty equivalent of the lottery. There appears to be some robustness, which suggests that lacking this information may not be too much of a problem, as long as loss-averse individuals are provided with insurance as derived above.

Finally, we have assumed throughout our analysis that the degree of loss aversion is commonly known. If, instead, we assumed that these parameters are private information, a hard multi-dimensional mechanism design problem arises. Our analysis nevertheless provides some insights into this problem. We could relax the assumption that the loss-aversion parameters in the money dimension are commonly known and allow them to be distributed arbitrarily, as the designer optimally eliminates any ex-post variation in the transfers irrespective of the degree of loss aversion. We leave the question of private information regarding the degree of loss aversion in the trade dimension for further research.

A Proofs

Proof of Proposition 1

Suppose f was CPEIC. Then, by definition the strategy profile s^t a CPE in the direct mechanism Γ^d and thus, again by definition, the direct mechanism implements f in CPE. Conversely, suppose there is a mechanism $\Gamma = (M_1, \ldots, M_N, g)$ that implements f in CPE. If $s^* = (s_1^*, \ldots, s_N^*)$ is a CPE, then for all $i, m_i' \in M_i$ and θ_i

$$U_i(s_i^*(\theta_i), s_{-i}^*, \Gamma|\theta_i) \geq U_i(m_i', s_{-i}^*, \Gamma|\theta_i)$$

by definition of the CPE. In particular, this is also true for $m_i' = s_i^*(\hat{\theta}_i)$ for all $i \in I, \hat{\theta}_i \in \Theta_i$. Therefore, given that $s^* = (s_1^*, \ldots, s_N^*)$ is a CPE we have for all $i \in I, \theta_i, \hat{\theta}_i \in \Theta_i$,

$$U_i(s_i^*(\theta_i), s_{-i}^*, \Gamma|\theta_i) \geq U_i(s_i^*(\hat{\theta}_i), s_{-i}^*, \Gamma|\theta_i)$$

Since Γ implements f in CPE we have

$$g(s_1^*(\theta_1), \ldots, s_N^*(\theta_N)) = f(\theta_1, \ldots, \theta_N),$$

implying

$$U_i(s_i^t(\theta_i), s_{-i}^t, \Gamma^d|\theta_i) \geq U_i(s_i^t(\hat{\theta}_i), s_{-i}^t, \Gamma^d|\theta_i)$$

for all $i \in I$, θ_i,$\hat{\theta}_i \in \Theta_i$. Thus, the truthful strategy profile s^t is a CPE in the direct mechanism and therefore the social choice function f is CPEIC.

Proof of Proposition 2

Proof. Suppose the social choice function f is CPEIC. Take some $\hat{\theta}_i > \theta_i$, then by CPEIC

$$U_i(\theta_i, s_{-i}^t|\theta_i) \geq \theta_i \tilde{v}_i(\hat{\theta}_i) + \tilde{t}_i(\hat{\theta}_i) = U_i(\hat{\theta}_i, s_{-i}^t|\hat{\theta}_i) + (\theta_i - \hat{\theta}_i)\tilde{v}_i(\hat{\theta}_i)$$

and analogously

$$U_i(\hat{\theta}_i, s_{-i}^t|\hat{\theta}_i) \geq \hat{\theta}_i \tilde{v}_i(\theta_i) + \tilde{t}_i(\theta_i) = U_i(\theta_i, s_{-i}^t|\theta_i) + (\hat{\theta}_i - \theta_i)\tilde{v}_i(\theta_i).$$

Thus,

$$\tilde{v}_i(\hat{\theta}_i) \geq \frac{U_i(\hat{\theta}_i, s_{-i}^t|\hat{\theta}_i) - U_i(\theta_i, s_{-i}^t|\theta_i)}{\hat{\theta}_i - \theta_i} \geq \tilde{v}_i(\theta_i),$$

implying that $\tilde{v}_i$ is non-decreasing because we assumed $\hat{\theta}_i > \theta_i$. Now, letting $\hat{\theta}_i \to \theta_i$ we get that for all θ_i we have

$$\frac{\partial U_i(\theta_i, s^t_{-i}|\theta_i)}{\partial \theta_i} = \tilde{v}_i(\theta_i)$$

and so

$$U_i(\theta_i, s^t_{-i}|\theta_i) = U_i(0, s^t_{-i}|0) + \int_0^{\theta_i} \tilde{v}_i(s)\, ds$$

for all $\theta_i \in \Theta_i$. Conversely, suppose that conditions (i) and (ii) hold. Without loss of generality, take any $\theta_i > \hat{\theta}_i$. Then,

$$\begin{aligned} U_i(\theta_i, s^t_{-i}|\theta_i) - U_i(\hat{\theta}_i, s^t_{-i}|\hat{\theta}_i) &= \int_{\hat{\theta}_i}^{\theta_i} \tilde{v}_i(s)\, ds \\ &\geq \int_{\hat{\theta}_i}^{\theta_i} \tilde{v}_i(\hat{\theta}_i)\, ds \\ &= (\theta_i - \hat{\theta}_i)\tilde{v}_i(\hat{\theta}_i). \end{aligned}$$

Hence,

$$U_i(\theta_i, s^t_{-i}|\theta_i) \geq U_i(\hat{\theta}_i, s^t_{-i}|\hat{\theta}_i) + (\theta_i - \hat{\theta}_i)\tilde{v}_i(\hat{\theta}_i) = \theta_i \tilde{v}_i(\hat{\theta}_i) + \tilde{t}_i(\hat{\theta}_i)$$

and similarly

$$U_i(\hat{\theta}_i, s^t_{-i}|\hat{\theta}_i) \geq U_i(\theta_i, s^t_{-i}|\theta_i) + (\hat{\theta}_i - \theta_i)\tilde{v}_i(\theta_i) = \hat{\theta}_i \tilde{v}_i(\theta_i) + \tilde{t}_i(\theta_i).$$

Consequently, f is CPEIC. ■

Proof of Proposition 3

As noted in the main text we can write $\tilde{v}_B(\theta_B) = y_B(\theta_B)(1 - \Lambda_B(1 - y_B(\theta_B)))$ and $\tilde{v}_S(\theta_S) = y_S(\theta_S)(1+\Lambda_S(1-y_S(\theta_S)))$, allowing us to rewrite expected utility as in equations (4) and (5) to

$$\begin{aligned} U_S(\theta_S, s^t_B|\theta_S) &= U_S(b_S, s^t_B|b_S) + \int_{\theta_S}^{b_S} y_S(t)(1 + \Lambda_S(1 - y_S(t)))\, dt, \\ U_B(\theta_B, s^t_S|\theta_B) &= U_B(a_B, s^t_S|a_B) + \int_{a_B}^{\theta_B} y_B(t)(1 - \Lambda_B(1 - y_B(t)))\, dt. \end{aligned}$$

Taking derivatives we obtain

$$\frac{\partial U_S(\theta_S, s_B^t|\theta_S)}{\partial \Lambda_S} = \int_{\theta_S}^{b_S} y_S(t)(1 - y_S(t))\, dt \geq 0$$

$$\frac{\partial U_B(\theta_B, s_S^t|\theta_B)}{\partial \Lambda_B} = -\int_{a_B}^{\theta_B} y_B(t)(1 - y_B(t))\, dt \leq 0$$

Proof of Proposition 4

We begin by noting that

$$\begin{aligned}
&\tilde{v}_B(\theta_B) \\
&= \int_{a_S}^{b_S} y^f(\theta_S, \theta_B)\, dF_S(\theta_S) + \eta_B^1 \int_{a_S}^{b_S} \int_{a_S}^{b_S} \mu_B^1 \left(y^f(\theta_S, \theta_B) - y^f(\theta'_S, \theta_B)\right)\, dF_S(\theta'_S)\, dF_S(\theta_S), \\
&= y_B(\theta_B) + \eta_B^1 \int_{a_S}^{b_S} \int_{a_S}^{b_S} y^f(\theta_S, \theta_B)(1 - y^f(\theta'_S, \theta_B)) - \lambda_B^1 (1 - y^f(\theta_S, \theta_B)) y^f(\theta'_S, \theta_B)\, dF_S(\theta'_S)\, dF_S(\theta_S), \\
&= y_B(\theta_B)(1 + \Lambda_B(y_B(\theta_B) - 1))
\end{aligned}$$

and analogously $\tilde{v}_S(\theta_S) = y_S(\theta_S)(1 - \Lambda_S(y_S(\theta_S) - 1))$, where

$$y_B(\theta_B) = \int_{a_S}^{b_S} y^f(\theta_S, \theta_B) dF_S(\theta_S), \quad y_S(\theta_S) = \int_{a_B}^{b_B} y^f(\theta_S, \theta_B) dF_B(\theta_B).$$

Imposing CPEIC we can write the sum of the agents' ex ante expected utilities as

$$\begin{aligned}
&\int_{a_B}^{b_B} U_B(\theta_B) f_B(\theta_B) d\theta_B + \int_{a_S}^{b_S} U_S(\theta_S) f_S(\theta_S) d\theta_S \\
&= U_B(a_B) + \int_{a_B}^{b_B} \int_{a_B}^{\theta_B} y_B(t)(1 + \Lambda_B(y_B(t) - 1)) dt f_B(\theta_B) d\theta_B \\
&+ U_S(b_S) + \int_{a_S}^{b_S} \int_{\theta_S}^{b_S} y_S(t)(1 - \Lambda_S(y_S(t) - 1)) dt f_S(\theta_S) d\theta_S \\
&= U_B(a_B) + \int_{a_B}^{b_B} y_B(\theta_B)(1 + \Lambda_B(y_B(\theta_B) - 1))(1 - F_B(\theta_B)) d\theta_B \\
&+ U_S(b_S) + \int_{a_S}^{b_S} y_S(\theta_S)(1 - \Lambda_S(y_S(\theta_S) - 1)) F_S(\theta_S) d\theta_S.
\end{aligned}$$

Note that the monotonicity constraints are satisfied due to Assumption 1, i.e., $\Lambda_B, \Lambda_S \leq 1$. Further, from the discussion in the main text we know that we can set the loss aversion in the money dimension to zero, as it only makes the problem harder. This allows us to express the sum of the agents' ex ante expected utilities as

$$\int_{a_B}^{b_B} U_B(\theta_B) f_B(\theta_B) d\theta_B + \int_{a_S}^{b_S} U_S(\theta_S) f_S(\theta_S) d\theta_S$$

$$= \int_{a_B}^{b_B} \int_{a_S}^{b_S} (\theta_B - \theta_S) y(\theta_S, \theta_B) f_S(\theta_S) f_B(\theta_B) d\theta_S d\theta_B$$
$$+ \int_{a_S}^{b_S} \theta_S y_S(\theta_S) \Lambda_S(y_S(\theta_S) - 1) f_S(\theta_S) d\theta_S + \int_{a_B}^{b_B} \theta_B y_B(\theta_B) \Lambda_B(y_B(\theta_B) - 1) f_B(\theta_B) d\theta_B$$

where we used CPEIC and integration by parts towards the end. Putting these two equations together we get

$$U_B(a_B) + U_S(b_S)$$
$$= \int_{a_B}^{b_B} \int_{a_S}^{b_S} (\theta_B - \theta_S) y(\theta_S, \theta_B) f_S(\theta_S) f_B(\theta_B) d\theta_S d\theta_B$$
$$+ \int_{a_S}^{b_S} \theta_S y_S(\theta_S) \Lambda_S(y_S(\theta_S) - 1) f_S(\theta_S) d\theta_S + \int_{a_B}^{b_B} \theta_B y_B(\theta_B) \Lambda_B(y_B(\theta_B) - 1) f_B(\theta_B) d\theta_B$$
$$- \int_{a_B}^{b_B} y_B(\theta_B)(1 + \Lambda_B(y_B(\theta_B) - 1))(1 - F_B(\theta_B)) d\theta_B - \int_{a_S}^{b_S} y_S(\theta_S)(1 - \Lambda_S(y_S(\theta_S) - 1)) F_S(\theta_S) d\theta_S.$$

Individual rationality requires $U_B(a_B) + U_S(b_S) \geq 0$. We will now show that this condition is never satisfied for any combination of buyer and seller loss aversion. From our discussion in the main text, we know that it is sufficient to consider the case $\Lambda_S = 0$, i.e., no loss aversion on the trade-dimension for the seller. This allows us to simplify and rewrite to

$$U_B(a_B) + U_S(b_S)$$
$$= \int_{a_B}^{b_B} \int_{a_S}^{b_S} \left(\left[\theta_B - \frac{1 - F_B(\theta_B)}{f_B(\theta_B)} \right] - \left[\theta_S + \frac{F_S(\theta_S)}{f_S(\theta_S)} \right] \right) y(\theta_S, \theta_B) f_B(\theta_B) f_S(\theta_S) d\theta_S d\theta_B$$
$$+ \Lambda_B \int_{a_B}^{b_B} y_B(\theta_B)(y_B(\theta_B) - 1) \left[\theta_B - \frac{1 - F_B(\theta_B)}{f_B(\theta_B)} \right] f_B(\theta_B) d\theta_B.$$

Myerson and Satterthwaite (1983) show in their proof of Theorem 1 (p. 269) that

$$\int_{a_B}^{b_B} \int_{a_S}^{b_S} \left(\left[\theta_B - \frac{1 - F_B(\theta_B)}{f_B(\theta_B)} \right] - \left[\theta_S + \frac{F_S(\theta_S)}{f_S(\theta_S)} \right] \right) y(\theta_S, \theta_B) f_B(\theta_B) f_S(\theta_S) d\theta_S d\theta_B$$
$$= - \int_{a_B}^{b_S} (1 - F_B(x)) F_S(x) \, dx.$$

Further, we have $y_B(\theta_B) = F_S(\theta_B)$ since we are considering the ex-post efficient mechanism. Putting this together yields

$$U_B(a_B) + U_S(b_S) = - \int_{a_B}^{b_S} (1 - F_B(x)) F_S(x) \, dx$$
$$+ \Lambda_B \int_{a_B}^{b_B} F_S(x)(F_S(x) - 1) \left[x - \frac{1 - F_B(x)}{f_B(x)} \right] f_B(x) dx.$$

Careful inspection of the limits of the integrals shows that

$$
\begin{aligned}
&U_B(a_B) + U_S(b_S) = -\int_{\max\{a_B,a_S\}}^{\min\{b_S,b_B\}} (1 - F_B(x))F_S(x)\, dx \\
&+ \Lambda_B \int_{\max\{a_B,a_S\}}^{\min\{b_S,b_B\}} F_S(x)(F_S(x) - 1)\left[x - \frac{1 - F_B(x)}{f_B(x)}\right] f_B(x)dx \\
&= -\int_{\max\{a_B,a_S\}}^{\min\{b_S,b_B\}} (1 - F_B(x))F_S(x) + \Lambda_B(1 - F_S(x))F_S(x)\left[x - \frac{1 - F_B(x)}{f_B(x)}\right] f_B(x)dx \\
&= -\int_{\max\{a_B,a_S\}}^{\min\{b_S,b_B\}} (1 - F_B(x))F_S(x)(1 - \Lambda_B(1 - F_S(x))) + \Lambda_B(1 - F_S(x))F_S(x)x f_B(x)\, dx \\
&< 0,
\end{aligned}
$$

violating individual rationality. To conclude the proof, recall from our discussion of the information rents, that loss aversion in the money dimension makes the problem unambiguously harder, as it reduces the gains from trade without affecting the information rents. Thus, impossibility in the absence of loss aversion in the money dimension implies impossibility in the presence of loss aversion in the money dimension.

Proof of Proposition 5

As noted in the text, the objective function reads

$$
\begin{aligned}
&\int_a^b \left(\eta_B^2 w_B(\theta_B) + \theta_B \tilde{v}_B(\theta_B) - \int_a^{\theta_B} \tilde{v}_B(t)\, dt\right)\, dF_B(\theta_B) \\
&+ \int_a^b \left(\eta_S^2 w_S(\theta_S) - \theta_S \tilde{v}_S(\theta_S) - \int_{\theta_S}^b \tilde{v}_S(t)\, dt\right)\, dF_S(\theta_S),
\end{aligned}
$$

where we observe that w_B and w_S enter positively. Next, note that

$$
\begin{aligned}
w_S(\theta) &= \int_a^b \int_a^b \mu_S^2 \left(t_S^f(\theta_S,\theta) - t_S^f(\theta_S,\theta')\right)\, dF_B(\theta')\, dF_B(\theta) \\
&= \int_a^b \int_a^b \left(t_S^f(\theta_S,\theta) - t_S^f(\theta_S,\theta')\right) \mathbb{1}[t_S^f(\theta_S,\theta) > t_S^f(\theta_S,\theta')]\, dF_B(\theta')\, dF_B(\theta) \\
&+ \int_a^b \int_a^b \lambda_S^2 \left(t_S^f(\theta_S,\theta) - t_S^f(\theta_S,\theta')\right) \mathbb{1}[t_S^f(\theta_S,\theta) < t_S^f(\theta_S,\theta')]\, dF_B(\theta')\, dF_B(\theta) \\
&= \int_a^b \int_a^b \left(t_S^f(\theta_S,\theta) - t_S^f(\theta_S,\theta')\right) \mathbb{1}[t_S^f(\theta_S,\theta) > t_S^f(\theta_S,\theta')]\, dF_B(\theta')\, dF_B(\theta) \\
&- \lambda_S^2 \int_a^b \int_a^b \left(t_S^f(\theta_S,\theta') - t_S^f(\theta_S,\theta)\right) \mathbb{1}[t_S^f(\theta_S,\theta') > t_S^f(\theta_S,\theta)]\, dF_B(\theta')\, dF_B(\theta) \\
&= (1 - \lambda_S^2) \int_a^b \int_a^b \left(t_S^f(\theta_S,\theta') - t_S^f(\theta_S,\theta)\right) \mathbb{1}[t_S^f(\theta_S,\theta') > t_S^f(\theta_S,\theta)] dF_B(\theta') dF_B(\theta),
\end{aligned}
$$

where $\mathbb{1}$ denotes the indicator function. The key step in the above derivation lies in the last equality. Comparing the two integrands on the third and second-to-last lines, we notice

that they look the same but that θ_B and θ'_B are interchanged. To see the equality, change the order of integration in the integral on the second-to-last line and perform a change of variables for the resulting integral. This shows that the two integrals are actually the same and allows us to sum them. Thus, since $\lambda_S^2 > 1$ we find $w_S(\theta_S) \leq 0$. The argument for w_B is analogous.

Given that w_B and w_S enter the designer's objective function positively, the designer optimally sets $w_i(\theta_i) = 0$. Further, a transfer achieves $w_i(\theta_i) = 0$ if and only if the transfer is independent of almost all buyer types. Thus, interim deterministic transfers are the only transfers that achieve $w_i(\theta_i) = 0$.

Proof of Lemma 1

We begin by proving the following technical lemma.

Lemma 2 *Let $f, g : [a, b] \to \mathbb{R}$ be two integrable functions with the following properties:*

(1) We have

$$\int_a^b f(x)dx \geq \int_a^b g(x)dx$$

(2) There exists a $x_0 \in [a, b]$ such that

a) $f(x) \geq g(x)$ for a.e. $x \leq x_0$

b) $f(x) \leq g(x)$ for a.e. $x \geq x_0$

Further, define $\varphi : [a, b] \to \mathbb{R}$. Then, if φ is monotonically decreasing we have

$$\int_a^b \varphi(x)f(x)dx \geq \int_a^b \varphi(x)g(x)dx,$$

and if φ is monotonically increasing we have

$$\int_a^b \varphi(x)f(x)dx \leq \int_a^b \varphi(x)g(x)dx.$$

Proof. We prove the statement for the case when φ is monotonically decreasing. For the case of an increasing φ simply reverse the appropriate inequalities. We begin by rewriting property (1) to

$$\int_a^b f(x)dx \geq \int_a^b g(x)dx \Leftrightarrow \int_a^{x_0} (f(x) - g(x))dx \geq \int_{x_0}^b (g(x) - f(x))dx$$

and note that both integrands are weakly positive due to property (2). Then, once more by (2), we have for a.e. $x \leq x_0$

$$\varphi(x)(f(x) - g(x)) \geq \varphi(x_0)(f(x) - g(x))$$

for a.e. $x \leq x_0$, which we can integrate to obtain

$$\int_a^{x_0} \varphi(x)(f(x) - g(x))dx \geq \int_a^{x_0} \varphi(x_0)(f(x) - g(x))dx. \tag{14}$$

Proceeding analogously, we obtain the inequality

$$\int_{x_0}^b \varphi(x_0)(g(x) - f(x))dx \geq \int_{x_0}^b \varphi(x)(g(x) - f(x))dx. \tag{15}$$

Further, we also have by

$$\int_a^{x_0} \varphi(x_0)(f(x) - g(x))dx = \int_{x_0}^b \varphi(x_0)(g(x) - f(x))dx \tag{16}$$

by property (1). Combining the inequalities in equations (14) to (16) we obtain

$$\int_a^{x_0} \varphi(x)(f(x) - g(x))dx \geq \int_{x_0}^b \varphi(x)(g(x) - f(x))dx,$$

which we can rearrange to

$$\int_a^b \varphi(x)f(x)dx \geq \int_a^b \varphi(x)g(x)dx,$$

completing the proof. ■

With this in hand, we can prove Lemma 1. Let $\hat{q} : [0,1] \times [0,1] \to \{0,1\}$ be any candidate for optimality. Associate to $\hat{q}$ the function

$$q(v_B, v_S) = \begin{cases} 1 & 0 \leq v_S \leq \hat{q}_B(v_B) \\ 0 & o.w., \end{cases} \tag{17}$$

where $\hat{q}_B(v_B) = \int_0^1 \hat{q}(v_B, v_S)dv_S$. First, note that $q_B = \hat{q}_B$ by construction. Thus, the first integral in equation (8) is not affected by a change from $\hat{q}$ to q. However, we will show that the second integral, which enters negatively, will become smaller. To do so, we will show

$$\int_0^1 M_S(v_S)q_S(v_S)(1+\Lambda_S)dv_S \leq \int_0^1 M_S(v_S)\hat{q}_S(v_S)(1+\Lambda_S)dv_S \tag{18}$$

and

$$\int_0^1 \Lambda_S M_S(v_S)(q_S^2(v_S) - \hat{q}_S^2(v_S))dv_S \geq 0. \tag{19}$$

To prove (18), fix $v_B \in [0,1]$ and apply Lemma 2 by setting $f(v_S) = q(v_B, v_S)$, $g(v_S) = \hat{q}(v_B, v_S)$ and $x_0 = \hat{q}_B(v_B)$. Note that the properties (1) and (2) in the lemma are satisfied by the construction of q from $\hat{q}$. Further, $M_S(v_S)(1+\Lambda_S)$ is a monotonically increasing function so that Lemma 2 yields

$$\int_0^1 M_S(v_S)(1+\Lambda_S)q(v_B, v_S)dv_S \geq \int_0^1 M_S(v_S)(1+\Lambda_S)\hat{q}(v_B, v_S)dv_S.$$

Let us now integrate this with respect to v_B and apply Fubini's theorem to reverse the order of integration to obtain

$$\begin{aligned}
&\int_0^1 \int_0^1 M_S(v_S)(1+\Lambda_S)q(v_B, v_S)dv_S dv_B \geq \int_0^1 \int_0^1 M_S(v_S)(1+\Lambda_S)\hat{q}(v_B, v_S)dv_S dv_B \\
&\Leftrightarrow \int_0^1 \int_0^1 M_S(v_S)(1+\Lambda_S)q(v_B, v_S)dv_B dv_S \geq \int_0^1 \int_0^1 M_S(v_S)(1+\Lambda_S)\hat{q}(v_B, v_S)dv_B dv_S \\
&\Leftrightarrow \int_0^1 M_S(v_S)q_S(v_S)(1+\Lambda_S)dv_S \leq \int_0^1 M_S(v_S)\hat{q}_S(v_S)(1+\Lambda_S)dv_S
\end{aligned}$$

as claimed in equation (18).

To prove (19), we begin by noting that we can rewrite this inequality to

$$\begin{aligned}
&\int_0^1 M_S(v_S)[q_S^2(v_S) - \hat{q}_S^2(v_S)]dv_S \geq 0 \\
&\Leftrightarrow \int_0^1 [q_S(v_S) + \hat{q}_S(v_S)][M_S(v_S)q_S(v_S) - M_S(v_S)\hat{q}_S(v_S)]dv_S \geq 0.
\end{aligned}$$

We will once more apply Lemma 2. Fix $v_B \in [0,1]$ and set $f(v_S) = M_S(v_S)q(v_B, v_S)$, $g(v_S) = M_S(v_S)\hat{q}(v_B, v_S)$. Note that by (18)

$$\int_0^1 f(v_S)dv_S \geq \int_0^1 g(v_S)dv_S$$

so that property (1) is satisfied. Further, if $v_S \leq \hat{q}_B(v_B)$, then $M_S(v_S)q(v_S, v_B) = M_S(v_S) \geq M_S(v_S)\hat{q}(v_B, v_S)$. Similarly, if $v_S \geq \hat{q}_B(v_B)$, then $M_S(v_S)q(v_S, v_B) = 0 \leq M_S(v_S)\hat{q}(v_B, v_S)$. Together, this shows that property (2) is satisfied. Further, define

$\phi(v_S) = q_S(v_S) + \hat{q}_S(v_S)$ and note that it is a decreasing function, as the candidate solution $\hat{q}_S$ is decreasing bq assumption and the associated q_S by the construction in (17).[27] Therefore, it follows from Lemma 2 and by once more integrating with respect to v_B and applying Fubini's theorem that

$$\begin{aligned}
&\int_0^1 [q_S(v_S) + \hat{q}_S(v_S)][M_S(v_S)q_S(v_S) - M_S(v_S)\hat{q}_S(v_S)]dv_S \geq 0 \\
&\Leftrightarrow \int_0^1 [q_S(v_S) + \hat{q}_S(v_S)]M_S(v_S)(q_S(v_S) - \hat{q}_S(v_S))dv_S \geq 0 \\
&\Leftrightarrow \int_0^1 \Lambda_S M_S(v_S)(q_S^2(v_S) - \hat{q}_S^2(v_S))dv_S \geq 0,
\end{aligned}$$

as claimed in equation (19).

Putting equations (18) and (19) together, we obtain that

$$\begin{aligned}
&\int_0^1 M_S(v_S)\hat{q}_S(v_S)(1 - \Lambda_S(\hat{q}_S(v_S) - 1))\, dv_S \\
&- \int_0^1 M_S(v_S)q_S(v_S)(1 - \Lambda_S(q_S(v_S) - 1))\, dv_S \\
=&\int_0^1 M_S(v_S)(1 + \Lambda_S)\hat{q}_S(v_S) - M_S(v_S)\Lambda_S\hat{q}_S^2(v_S)\, dv_S \\
&- \int_0^1 M_S(v_S)(1 + \Lambda_S)q_S(v_S) - M_S(v_S)\Lambda_S q_S^2(v_S)\, dv_S \\
=&\int_0^1 M_S(v_S)(1 + \Lambda_S)[\hat{q}_S(v_S) - q_S(v_S)] + M_S(v_S)\Lambda_S[q_S^2(v_S) - \hat{q}_S^2(v_S)]\, dv_S \geq 0,
\end{aligned}$$

showing that the second integral in equation (8) indeed becomes smaller when moving from $\hat{q}$ to q. Hence, for any $\hat{q}$ not of the form (9) we can construct a function in this class which does better, completing the proof.

Proof of Corollary 2

Imposing that types are uniformly distributed types on the interval $[a, a+1]$ for $a \geq 0$, the condition in equation (13) can be written as

$$\begin{aligned}
&\frac{2\theta_B - a - 1}{1 - \Lambda_S + 2\Lambda_S(\theta_B - a)} \geq \frac{2\theta_S - a}{1 - \Lambda_B + 2\Lambda_B(\theta_S - a)} \\
&\theta_S \leq \delta(\theta_B) = \frac{(2\theta_B - 1 - a)(1 - \Lambda_B(2a+1) + a\Lambda_S) + a - \Lambda_S a^2}{2(1 - \Lambda_B(2\theta_B - a - 1) + \Lambda_S(2\theta_B - 1 - 2a))} =: \theta(\theta_B)
\end{aligned}$$

[27] To see this, note that $q_S(v_S) = 1 - q_B^{-1}(v_S)$. Thus, since q_B is increasing q_S is decreasing.

and the conditions on the degree of loss aversion reduce to $IC = \{\Lambda_B \leq 1/(a+1), \Lambda_S \leq \min\{1, 1/a\}\}$. From here one can show that the trade rule induces less trade for given Λ_B, Λ_S as a increases, eventually eliminating trade altogether.

Proof of Proposition 8

The derivations of the mechanisms maximizing the total and the material gains from trade proceed analogously. We here present the derivations for the case of maximizing the total gains from trade. Making use of Proposition 7 and the budget constraint (AB), we eliminate the transfers from the problem and can rewrite the objective function to

$$\begin{aligned}
&\int_a^b U_B(\theta_B, s_S^t|\theta_B)\, dF_B(\theta_B) + \int_a^b U_S(\theta_S, s_B^t|\theta_S)\, dF_S(\theta_S) \\
&= \int_a^b \left(\theta_B y_B(\theta_B)(1 + \Lambda_B(y_B(\theta_B) - 1)) - \bar{t}_B(\theta_B) + \eta_B^2 w_B(\theta_B)\right)\, dF_B(\theta_B) \\
&\quad - \int_a^b \left(\theta_S y_S(\theta_S)(1 - \Lambda_S(y_S(\theta_S) - 1)) - \bar{t}_S(\theta_S) - \eta_S^2 w_S(\theta_S)\right)\, dF_S(\theta_S) \\
&= \int_a^b \theta_B y_B(\theta_B)(1 + \Lambda_B(y_B(\theta_B) - 1))\, dF_B(\theta_B) - \int_a^b \theta_S y_S(\theta_S)(1 - \Lambda_S(y_S(\theta_S) - 1))\, dF_S(\theta_S).
\end{aligned}$$

Further, the budget constraint AB and the CPEIC can be jointly written as

$$\begin{aligned}
&\int_a^b J_B(\theta_B) y_B(\theta_B) \left(1 - \Lambda_B \left(1 - y_B(\theta_B)\right)\right)\, dF_B(\theta_B) \\
&= \int_a^b J_S(\theta_S) y_S(\theta_S) \left(1 + \Lambda_S \left(1 - y_S(\theta_S)\right)\right)\, dF_S(\theta_S),
\end{aligned}$$

as well as the monotonicity constraints. We can set up the Lagrangian as

$$\begin{aligned}
\mathcal{L}(y^f, \gamma) = &\int_a^b (\theta_B + \gamma J_B(\theta_B)) y_B(\theta_B) \left(1 - \Lambda_B \left(1 - y_B(\theta_B)\right)\right)\, dF_B(\theta_B) \\
&- \int_0^1 (\theta_S + \gamma J_S(\theta_S)) y_S(\theta_S) \left(1 + \Lambda_S \left(1 - y_S(\theta_S)\right)\right)\, dF_S(\theta_S).
\end{aligned}$$

Note that we must have $\gamma \geq 0$, because relaxing the budget constraint (i.e., allowing the designer to run a deficit) can only increase the objective. Hence, by Assumption 2, $\theta_B + \gamma J_B(\theta_B)$ and $\theta_S + \gamma J_S(\theta_S)$ are strictly increasing in θ_B and θ_S, respectively. Therefore, the arguments from the proof of the revenue maximizing mechanism carry through and we obtain

$$\begin{aligned}\mathcal{L}(y^f,\gamma) &= \int_a^b\int_a^b \underbrace{(\theta_B+\gamma J_B(\theta_B))\left[(1-\Lambda_B)+2\Lambda_B F_S(\theta_S)\right]}_{:=\tilde{J}_B^{TG}(\theta_B,\theta_S,\gamma)} y^f(\theta_B,\theta_S)dF_S(\theta_S)dF_B(\theta_B)\\ &\quad - \int_a^b\int_a^b \underbrace{(\theta_S+\gamma J_S(\theta_S))\left[(1+\Lambda_S)-2\Lambda_S(1-F_B(\theta_B))\right]}_{\tilde{J}_S^{TG}(\theta_S,\theta_B,\gamma)} y^f(\theta_B,\theta_S)dF_S(\theta_S)dF_B(\theta_B),\\ &= \int_a^b\int_a^b \left(\tilde{J}_B^{TG}(\theta_B,\theta_S,\gamma)-\tilde{J}_S^{TG}(\theta_S,\theta_B,\gamma)\right) y^f(\theta_B,\theta_S)dF_S(\theta_S)dF_B(\theta_B).\end{aligned}$$

We now have a concave maximization problem so that an trade rule y^f is optimal if and only if (see, e.g., Theorems 1 and 2 in Luenberger, 1969, p. 217 and p. 221)

$$y^f(\theta_B,\theta_S)=\begin{cases}1 & \text{if } \tilde{J}_B^{TG}(\theta_B,\theta_S,\gamma)-\tilde{J}_S^{TG}(\theta_S,\theta_B,\gamma)\geq 0\\ 0 & \text{otherwise}\end{cases} \tag{20}$$

subject to the monotonicty constraints on y_B and y_S. Leveraging Assumption 1 we can reformulate this to

$$\bar{J}_B^{TG}(\theta_B,\gamma) := \frac{\theta_B+\gamma J_B(\theta_B)}{1-\Lambda_S+2\Lambda_S F_B(\theta_B)} \geq \frac{\theta_S+\gamma J_S(\theta_S)}{1-\Lambda_B+2\Lambda_B F_S(\theta_S)} =: \bar{J}_S^{TG}(\theta_S,\gamma).$$

Finally, we define the set

$$IC^{TG}=\left\{(\Lambda_B,\Lambda_S)\geq 0 \mid \Lambda_j < \frac{1+\gamma J_i'(\theta_i)}{2(\theta_i+\gamma J_i(\theta_i))f_i(\theta_i)+(1+\gamma J_i'(\theta_i))(1-2F_i(\theta_i))}\ \forall\theta_i\in[a,b]\right\}.$$

The Lagrangian for the case of material gains from trade is obtained analogously and reads

$$\begin{aligned}\mathcal{L}(y^f,\gamma) &= \int_a^b\int_a^b \underbrace{(\theta_B+\gamma J_B(\theta_B)\left[(1-\Lambda_B)+2\Lambda_B F_S(\theta_S)\right])}_{:=\tilde{J}_B^{MG}(\theta_B,\theta_S,\gamma)} y^f(\theta_B,\theta_S)dF_S(\theta_S)dF_B(\theta_B)\\ &\quad - \int_a^b\int_a^b \underbrace{(\theta_S+\gamma J_S(\theta_S)\left[(1+\Lambda_S)-2\Lambda_S(1-F_B(\theta_B))\right])}_{\tilde{J}_S^{MG}(\theta_S,\theta_B,\gamma)} y^f(\theta_B,\theta_S)dF_S(\theta_S)dF_B(\theta_B),\\ &= \int_a^b\int_a^b \left(\tilde{J}_B^{MG}(\theta_B,\theta_S,\gamma)-\tilde{J}_S^{MG}(\theta_S,\theta_B,\gamma)\right) y^f(\theta_B,\theta_S)dF_S(\theta_S)dF_B(\theta_B).\end{aligned}$$

Proof of Corollary 3

To obtain the expression in Corollary 3 plug in the assumptions on the distributions of types and the parameters of loss aversion to rewrite the trade condition in equation (20)

to

$$\theta_S \leq \frac{(\theta_B + \gamma(2\theta_B - 1))(1 - \Lambda)}{1 + 2g - \Lambda}.$$

We can plug this into the budget constraint and solve for the Lagrange multiplier, yielding $\gamma = (\Lambda + \sqrt{1 + \Lambda + \Lambda^2})/2$ which yields the optimal trade rule

$$y^{TG}(\theta_B, \theta_S) = \begin{cases} 1 & \theta_S \leq \frac{(1-\Lambda)\left(2\theta_B\left(\sqrt{\Lambda^2+\Lambda+1}+\Lambda+1\right)-\sqrt{\Lambda^2+\Lambda+1}-\Lambda\right)}{2\left(\sqrt{\Lambda^2+\Lambda+1}+1\right)} \\ 0 & o.w. \end{cases}$$

Proceeding analogously for the case of material gains from trade, we obtain the same trade rule.

References

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011): "Reference Points and Effort Provision," *American Economic Review*, 101, 470–492.

Apesteguia, J. and M. Ballester (2015): "A Measure of Rationality and Welfare," *Journal of Political Economy*, 123, 1278–1310.

Bartling, B., L. Brandes, and D. Schunk (2015): "Expectations as Reference Points: Field Evidence from Professional Soccer," *Management Science*, 61, 2646–2661.

Bell, D. E. (1985): "Disappointment in Decision Making under Uncertainty," *Operations Research*, 33, 1–27.

Benkert, J.-M. (2022): "On the equivalence of optimal mechanisms with loss and disappointment aversion," *Economics Letters*, 214, mimeo.

Benkert, J.-M. and N. Netzer (2018): "Informational Requirements of Nudging," *Journal of Political Economy*, 126, 2323–2355.

Bernheim, B. and A. Rangel (2009): "Beyond Revealed Preference: Choice-Theoretic Foundations For Behavioral Welfare Economics," *Quarterly Journal of Economics*, 124, 51–104.

Bierbrauer, F. and N. Netzer (2016): "Mechanism Design and Intentions," *Journal of Economic Theory*, 163, 557–603.

Brown, A. L., T. Imai, F. M. Vieider, and C. F. Camerer (2021): "Meta-Analysis of EmpiricalEstimates of Loss-Aversion," Tech. rep., CESifo Working Papers.

Carbajal, J. C. and J. C. Ely (2016): "A Model of Price Discrimination under Loss Aversion and State-Contingent Reference Points," *Theoretical Economics*, 11, 455–485.

Chatterjee, K. and W. Samuelson (1983): "Bargaining under Incomplete Information," *Operations Research*, 31, 835–851.

Cramton, P., R. Gibbons, and P. Klemperer (1987): "Dissolving a partnership Efficienty," *Econometrica*, 55, 615–632.

Crawford, V. P. (2021): "Efficient Mechanisms for Level-k Bilateral Trading," *Games and Economic Behavior*, 127, 80–101.

Crawford, V. P. and J. Meng (2011): "New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income," *American Economic Review*, 101, 1912–1932.

Dato, S., A. Grunewald, D. Müller, and P. Strack (2017): "xpectation-basedlossaversionandstrategicinteraction," *Games and Economic Behavior*, 104, 681–705.

de Meza, D. and D. C. Webb (2007): "Incentive Design under Loss Aversion," *Journal of the European Economic Association*, 5, 66–92.

Dreyfuss, B., O. Heffetz, and M. Rabin (2019): "Expectations-Based Loss Aversion May Help Explain Seemingly Dominated Choices in Strategy-Proof Mechanisms," Tech. rep.

Driesen, B., A. Perea, and H. Peters (2012): "Alternating offers Bargaining with loss aversion," *Mathematical Social Sciences*, 64, 103–118.

Duraj, J. (2015): "Mechanism Design with News Utility," Personal Communcation.

——— (2018): "Mechanism Design with News Utility," Mimeo.

Eisenhuth, R. (2019): "Reference-Dependent Mechanism Design," *Economic Theory Bulletin*, 7, 77–103.

Eisenhuth, R. and M. Grunewald (2018): "Auctions with Loss Averse Bidders," *International Journal of Economic Theory*, 16, 129–152.

Ericson, K. M. M. and A. Fuster (2011): "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments," *Quarterly Journal of Economics*, 126, 1879–1907.

——— (2014): "The Endowment Effect," *Annual Review of Economics*, 6, 555–579.

Fehr, E. and L. Goette (2007): "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment," *American Economic Review*, 97, 298–317.

Fieseler, K., T. Kittsteiner, and B. Moldovanu (2003): "Partnerships, lemons, and efficient trade," *Journal of Economic Theory*, 113, 223–234.

Garratt, R. and M. Pycia (2020): "Efficient Bilateral Trade," Mimeo, UCSB and UZH.

Gershkov, A., B. Moldovanu, P. Strack, and M. Zhang (2021): "Optimal Auctions: Non-expected Utility and Constant Risk Aversion," *The Review of Economic Studies*.

Gill, D. and V. Prowse (2012): "A Structural Analysis of Disappointment Aversion in a Real Effort Competition," *American Economic Review*, 102, 469–503.

Gneezy, U., L. Goette, C. Sprenger, and F. Zimmermann (2017): "The Limits of Expectations-Based Reference Dependence," *Journal of the European Economic Association*, 15, 861–876.

Heffetz, O. (2021): "Are reference points merely lagged beliefs over probabilities?" *Journal of Economic Behavior and Organization*, 181, 252–269.

Heffetz, O. and J. A. List (2014): "Is the Endowment Effect an Expectations Effect?" *Journal of the European Economic Association*, 12, 1396–1422.

Heidhues, P. and B. Kőszegi (2014): "Regular Prices and Sales," *Theoretical Economics*, 9, 217–251.

Herweg, F., D. Müller, and P. Weinschenk (2010): "Binary Payment Schemes: Moral Hazard and Loss Aversion," *American Economic Review*, 100, 2451–2477.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–291.

Karle, H., G. Kirchsteiger, and M. Peitz (2015): "Loss Aversion and Consumption Choice: Theory and Experimental Evidence," *American Economic Journal: Microeconomics*, 7, 101–120.

Karle, H. and M. Möller (2020): "Selling in Advance to Loss Averse Consumers," *International Economic Review*, 61, 441–468.

Karle, H. and M. Peitz (2014): "Competition under consumer loss aversion," *The RAND Journal of Economics*, 45, 1–31.

Karle, H. and H. Schumacher (2017): "Advertising and Attachment: Exploiting Loss Aversion through Pre-Purchase Information," *RAND Journal of Economics*, 48, 875–1135.

Kőszegi, B. (2014): "Behavioral Contract Theory," *Journal of Economic Literature*, 52, 1075–1118.

Kőszegi, B. and M. Rabin (2006): "A Model of Reference-Dependent Preferences," *The Quarterly Journal of Economics*, 121, 1133–1165.

——— (2007): "Reference-Dependent Risk Attitudes," *The American Economic Review*, 97, 1047–1073.

——— (2009): "Reference-Dependent Consumption Plans," *American Economic Review*, 99, 909–936.

Kucuksenel, S. (2012): "Behavioral Mechanism Design," *Journal of Public Economic Theory*, 14, 767–789.

Loomes, G. and R. Sugden (1986): "Disappointment and Dynamic Consistency in Choice under Uncertainty," *The Review of Economic Studies*, 53, 271–282.

Manzini, P. and M. Mariotti (2014): "Welfare economics and bounded rationality: the case for model-based approaches," *Journal of Economic Methodology*, 21, 343–360.

Masatlioglu, Y. and C. Raymond (2016): "A Behavioral Analysis of Stochastic Reference Dependence," *The American Economic Review*, 106, 2760–2782.

Maskin, E. and J. Riley (1984): "Optimal Auctions with Risk Averse Buyers," *Econometrica*, 52, 1473 – 1518.

Meisner, V. and J. von Wangenheim (2021): "School choice and loss aversion," Tech. rep.

Myerson, R. B. (1981): "Optimal Auction Design," *Mathematics of Operations Research*, 6, 58.

Myerson, R. B. and M. A. Satterthwaite (1983): "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory*, 29, 265 – 281.

Pope, D. G. and M. E. Schweitzer (2011): "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes," *American Economic Review*, 101, 129–157.

Post, T., M. J. van den Assem, G. Baltussen, and R. H. Thaler (2008): "Dear or No Deal? Decision Making under Risk in a Large-Payoff Game Show," *American Economic Review*, 98, 38–71.

Rosato, A. (2016): "Selling Substitute Goods to Loss-Averse Consumers: Limited Availability, Bargains and Rip-offs," *Rand Journal of Economics*.

——— (2017): "Sequential Negotiations with Loss-Averse Buyers," *European Economic Review*, 91, 290–304.

——— (2021): "Loss aversion in sequential auctions," Working paper, University of Technology Sydney.

Rubinstein, A. and Y. Salant (2012): "Eliciting Welfare Preferences from Behavioural Data Sets," *Review of Economic Studies*, 79, 375–387.

Shalev, J. (2002): "Loss Aversion and Bargaining," *Theory and Decisions*, 52, 201–232.

Thaler, R. H. (1980): "Toward a positive theory of consumer choice," *Journal of Economic Behavior and Organization*, 1, 39–60.

——— (1999): "Mental Accounting Matters," *Journal of Behavioral Decision Making*, 12, 183–206.

Wolitzky, A. (2016): "Mechanism Design with Maxmin Agents: Theory and an Application to Bilateral Trade," *Theoretical Economics*, 11, 971–1004.