MAX PLANCK INSTITUTE
FOR RESEARCH ON
COLLECTIVE GOODS

CHRISTOPH ENGEL

Discussion Paper
2024/19

THE NEGOTIATION TRAP: AN EXPERIMENT ON A LARGE LANGUAGE MODEL

The Negotiation Trap

An Experiment on a Large Language Model*

Christoph Engel

Abstract

In an experiment on the large language model GPT-4o, a supplier always makes a higher profit if it replaces uniform contract terms with a set of terms between which the customer may choose. The extra profit results from price discrimination. There is a first order and a second order effect. The first order effect results from heterogeneous willingness to pay for a more protective term. The second order effect results from the possibility that contract choice is a signal for general willingness to pay for the traded commodity. In the experiment, the effect is bigger if the least protective version is labelled as the default, and more protective terms as an "upgrade". The effect is smaller if, conversely, the most protective version is labelled as the default and less protective (and cheaper) versions as an opportunity for "savings". The effect is also bigger if the supplier only sets the price after it knows which version of the contract the consumer chooses. The profit increasing effect of giving the consumer a choice is strong. There is no piece of demographic information that has a stronger effect. Most pieces of demographic information (which the supplier might, for instance, learn through cookie data) have a significantly smaller effect on profit. If the supplier combines cookie information about demographic markers with contract choice, it always makes an extra profit.

Keywords: forced choice of contract clause; price discrimination; large language model; experiment

JEL: C91, D01, D02, D12, D42, D91, K12

* Helpful comments by Stefan Bechtold and Eyal Zamir, as well as audiences at ETH Zurich and at the Center for Regulation and Contract Law at University Zurich are gratefully acknowledged.

1. Introduction

Research question. Boilerplate is good for suppliers and bad for consumers. The "theory of blanket assent" (Llewellyn 1960, 370) is a scam. Jurisdictions have been differently sensitive to the concern. German law is particularly protective. §§ 305 ff. BGB impose a long list of constraints to the design of a standard form contract.[1] The European Union has passed a directive that obliges member states to implement a somewhat less intrusive, but still impressive list of constraints.[2] In the US, the courts have stepped in. They control boilerplate (for an overview see Zamir and Ayres 2020, 320 f.), chiefly with the help of the doctrine of unconscionability (Leff 1967, Kornhauser 1976, Korobkin 2003),[3] and the doctrine of reasonable expectations (Thomas 1998, Hillman and Rachlinski 2002, 459f.).

Regulators and courts have chiefly been concerned that suppliers might abuse freedom of contract to impose individual contract terms on consumers that are unfair (EU Directive 93/13/EC, Recital 4). This is why certain clauses are prohibited altogether (e.g. Annex to EU Directive 93/13/EC), and other clauses are subject to an ad hoc judiciary assessment of their substantive fairness (e.g. § 308 BGB). These statutory instruments typically also regulate procedure, but only as a boundary condition: what is required for considering a term to be unilaterally imposed, and therefore subject to regulatory oversight? Conversely, when is a clause considered negotiated, and therefore justified under general freedom of contract (e.g. Art. 3 EU Directive 93/13/EC; § 305b BGB)? Hence in the perspective of regulators, the fact that consumers have had a say on the content of a clause is never considered a normative problem, and possibly even the reason why regulatory oversight is suspended.[4]

Yet arguably consumers are not only treated unfairly if a risk materializes during the implementation of the contract, and the contract unilaterally imposes the harm or cost on the buyer. The unfairness may also result from the fact that the seller appropriates a larger share of gains from trade. This is possible if the supplier leverages contract design to widen the scope for price discrimination. Conceivably there are two channels: consumers may have a differently pronounced willingness to pay for a more protective clause. Moreover the fact that the consumer chooses a more (or less) protective clause may tell the supplier something about the willingness to pay of this customer for the product or service itself. In this paper, I not only draw conceptual attention to this point. I also show empirically that the resulting potential for consumer exploitation is substantial.

Normative assessment. Whether extended price discrimination is a reason for regulatory intervention depends on the normative assessment. Consumers who care about particularly protective terms get what they want. Consumers who do not care (sufficiently) about the term in question get a discount, at least in comparison with the price they would have to pay with stronger protection. But as always with price discrimination, higher efficiency comes at a distributional cost: the closer the differentiated term–price scheme captures consumers' preferences, the more the provider can exploit them. The distributional effect is even stronger if the

1 English version available at https://www.gesetze-im-internet.de/englisch_bgb/englisch_bgb.pdf.
2 Council Directive 93/13/EC of 5 April 1993 on unfair terms in consumer contracts, OJ 1993 L 95/29.
3 Restatement (Second) of Contracts 211(3); Uniform Commercial Code sec. 2-302.
4 I will come back to this possibility in the discussion ***.

choice of a specific contract term is predictive for (other determinants of) willingness to pay. This is what the title of this paper is meant to capture: what originally appears like provider complacency can actually turn into a trap for the consumer.

If the consumer correctly anticipates the effect, she will only accept if the term-price pair that she wants is too good a deal to deny. Then the choice of terms works as a screening device, in the technical sense of mechanism design. But behaviorally, the effect may well be even bigger. Many consumers may not even notice price discrimination. Even if they do, it may not occur to them that their choice of term has had an effect on price that transcends the announced discount / mark-up. Even more so if, as routinely in online trade, the provider does not make list prices public, so that the individual consumer only sees the price offered to her individually. This is what the title of this paper is meant to capture: what originally appears like provider complacency can actually turn into a trap for the consumer.

Practical relevance. It is common wisdom that the negotiation exception in the rules on standard form contracts has little practical relevance, at least for B2C trade. This perception is intuitive. Consumers do not routinely go to stores, or to online sites, and hassle over clauses in lengthy contracts. As a matter of fact, very few consumers bother to read the terms of a contract even if they have to actively declare they did (Bakos, Marotta-Wurgler et al. 2014). Is the possibility for enhanced price discrimination via choice of contract clauses therefore purely academic? Not quite. Amazon optionally offers guaranteed delivery by a defined date.[5] Apple offers additional protection with its Apple Care package,[6] same as many car dealers,[7] software as a service providers,[8] or consumer finance providers offering protection against identity theft.[9]

For two correlated reasons in the not so distant future, such practices might become even more common. Production processes and sales are increasingly digitized. Providers are therefore more likely to have microdata from which they can infer the additional cost of a more consumer-friendly contract term, and the additional savings from a more business-friendly term. It becomes feasible to replace the one-size-fits-all set of terms with a portfolio of terms that come with a specific price tag. On the other hand, if trade is online, it may be easier for the firm to exploit this information for price discrimination, and harder for the consumer to notice that she is discriminated against. Consequently, in a thoroughly digitized business, offering differently protective terms may be a win-win-proposition for the firm. It attracts additional customers who care, and it makes a higher profit.

Giving consumers a choice between alternative terms, at a different price, is also beneficial for suppliers from a behavioral angle. It has been shown that rarely consumers even look at standard form contracts (Bakos, Marotta-Wurgler et al. 2014). Consequently suppliers have little

5 https://www.amazon.com/gp/help/customer/display.html?nodeId=202075470.
6 https://www.apple.com/de/support/products/.
7 https://www.ala.co.uk/insights/warranty/coverage/dealership-warranty#:~:text=Dealership%20warranties%20often%20come%20with,included%20in%20your%20car%20financing.
8 https://www.cloudeagle.ai/blogs/service-level-agreements.
9 https://www.consumerfinance.gov/ask-cfpb/what-is-identity-monitoring-or-identity-theft-service-en-1369/#:~:text=Identity%20theft%20services%20monitor%20personally,places%20for%20any%20unusual%20activity.

reason to expect that, through the design of the contract, they can induce consumer behavior that they deem profitable. Yet if each version of a clause comes with a different price tag, in the spirit of "forced choice" (Sunstein 2015), consumers cannot avoid making up their mind. This makes a choice between differently protective clauses a strong diagnostic tool.

Empirical strategy. These considerations show: it could be consistently explained why providers may be willing to negotiate terms, why consumers may have a hard time resisting, and why this may give providers a higher profit, at least partly at the cost of consumers. But is the concern real? Should consumers, and consumer advocates for that matter, worry about falling into the negotiation trap? This is an empirical question.

Essentially, the story developed in this introduction is a causal claim: the provider makes a higher profit if it offers a portfolio of contract terms, compared with offering the one term that maximizes profit when standardizing the term for all consumers. There are two claims about mechanism: one local, and one global. The local claim says: discriminating contract terms increases supplier profit if and because customers have heterogenous preferences over contract protection. The global claim says: the choice of (some) contract terms is correlated with willingness to pay for the product, and hence is an informative signal.

It is a good heuristic for testing a causal claim to ask: what would the ideal experiment look like (Angrist and Pischke 2008, 3)? For testing the main claim, one would need to randomly give some suppliers the possibility to offer a portfolio of term-price pairs, while others are constrained to pick a single pair. One would measure supplier profit (and both prices and quantity as the elements from which profit results). For making results meaningful, all suppliers would need to offer the same product, to the same group of potential customers. As competition would create dependence, and thereby destroy the possibility to estimate a population effect, one would need a multitude of otherwise identical markets, and each provider not facing any competition (or the same degree of competition from other suppliers whose behavior one does not use for the experiment). Obviously this is impossible, or at least impractical.

For testing the claims about mechanism, one would need a heterogeneous group of potential customers with known preferences. For testing the local mechanism, these preferences would have to exclusively differ with respect to the provision of the contract that is manipulated. For testing the global mechanism, preferences would also have to differ in respects that transcend the manipulated contract clause. For comparability, again each provider would have to meet an identical (heterogeneous) population of potential customers, but not be exposed to competition by other participants in the experiment. Clearly this version of the experiment is even less practical.

Weaker substitutes are conceivable. Very likely firms have engaged in A/B testing before offering additional layers of contractual protection. While in the spirit of a randomized control trial, such tests come with a host of challenges for statistical inference (nicely summarized by Larsen, Stallrich et al. 2024). An obvious challenge is dependence: each provider can only experiment with its own offers. Yet most importantly from the regulatory perspective adopted in this paper: providers keep the design and the results of their A/B tests confidential. In a vignette study, one might ask participants how they would react when given the option to select

a more protective contract clause. One could compare these choices with a baseline in which this option is not available. This empirical strategy comes with its own limitations (Atzmüller and Steiner 2010, Aguinis and Bradley 2014). Most importantly for the present context: one could only ask (potential) buyers, not a population of sellers.[10] One would therefore not learn what is normatively most relevant: whether and in which ways suppliers would exploit the scope for price discrimination. Another constraint is practical: one could at best test a very small number of alternative scenarios.

Given these limitations of alternative methods, large language models provide researchers with an intriguing complementary empirical tool. The advantages are two-fold: LLMs might make phenomena observable that could not be observed *in vivo*. With the help of an LLM one can test an unprecedented number of conditions. This option is particularly appealing for understanding the power of moderators.

Two features of LLMs are critical for their power as a supplementary tool for empirical legal research: Large language models have been trained on more human utterances than any human being will ever hear or read in her lifetime. One can prompt the model to estimate how human individuals would have reacted when faced with the same choice. There is of course no guarantee. But in another (also law related) context, I have shown that such a belief prompt strongly increases the alignment with human subjects who have been asked the same question (Engel and McAdams 2024, 268-271).

Moreoever one can access the LLM through an application programming interface (API). When accessing the LLM on this path, by default the process has no memory[11]. Hence later draws are not contaminated by former responses. In terms of statistical analysis, each response generates one independent observation. The researcher does not train the model to give the responses insinuated by the design of the experiment. Specifically, as in an experiment with human subjects, one can give every new instance of the LLM (in the same treatment) the same task. Asking repeatedly is meaningful if one sets “temperature” to a sufficiently high value (1 in my experiments). Then one generates an entire distribution of responses, reflecting the distribution of responses the LLM expects the human subjects to give that it is prompted to predict (for technical background see He, Zhang et al. 2018).

Experiments with LLMs are of course not perfect. As with lab experiments with human participants, one will eventually put more trust on treatment effects than on absolute measures. Even if the language model is more sensitive to the stimulus than typical human participants would (or less sensitive, for that matter), the way how the language model responds to alternative versions of the stimulus arguably points the researcher into the right direction.

Relying on this new (imperfect) empirical method is all the more warranted if the law acts according to the principle of precaution (Foster, Vecchia et al. 2000, Kriebel, Tickner et al. 2001). Even if a normative concern has not been proven with ultimate certainty, the law considers which one would be worse: that the law overshoots, because the seeming normative

10 One could only ask non-sellers how they believe sellers to react.
11 https://github.com/openai/openai-cookbook/issues/275#issuecomment-1539239709.

concern is overstated; or that the law abstains although a serious concern would have called for intervention? Arguably regulatory responses to abusive contract design are a worthy candidate for the precautionary principle. The regulator intends to mitigate a power imbalance resulting from economies of scale. While the individual buyer meets this supplier only once, the seller interacts with a whole community of buyers, and hence can afford much higher transaction cost in drafting the contract. This suggests that, for motivating regulatory intervention, a reasonably substantiated risk of abuse suffices. Consequently, even taking into account that LLM evidence cannot be regarded as conclusive, regulators and courts might still consider clear LLM results to be a sufficient reason for intervention.

Present experiment. To assess whether the risk of a negotiation trap is to be taken seriously, I have written the sketch of case with the following properties: a) if it may employ a contract term in a standard form contract, the provider would save cost; b) the provider has the possibility to give the customer a choice between three different version of the relevant contract term, holding the product and all other contract terms constant; c) in the baseline the provider refrains from relying on the option. In the first treatment, it adds the option, but has no further information about the customer. I additionally manipulate whether the provider announces price upfront, or only after learning the customer's choice, and whether the option is only offered as such, or whether it is framed as either an "upgrade" (from the least protective version) or as the possibility of "savings", compared with the most protective, but also most expensive version. In the second treatment, I compare the effect of the option with demographic information purportedly signalling a higher or lower willingness to pay for the product. In the third treatment, I compare just having access to the respective piece of demographic information with additionally employing the option.

Preview of results. GPT comes to the conclusion that suppliers make at least a 15% higher profit when offering three alternative contract terms, rather than a uniform contract. If they only set their price after the consumer has indicated which contract clause she prefers, they even make a 39% higher profit. This profit is 8% higher than if the supplier commits upfront to a price for each version of the contract. The profit is 14% higher than without the option if the option is framed as an "upgrade", compared with the possibility for "savings". Offering alternative versions of the single clause increases profit as strongly as cookie information about the customer living in a wealthy neighbourhood, or having been on five intercontinental business class flights last year. For each of 10 different pieces of demographic information (again purportedly gleaned from cookie data), offering the three alternative versions of the one contract clause significantly increases profit.

Organisation of the paper. In the next section, the paper is positioned in the literature. Section 3 explains in which ways a portfolio of contract terms might create scope for price discrimination, and formulates the hypotheses to be tested. Section 4 introduces the design of the experiment. Section 5 reports results. Section 6 concludes with discussion.

2. Literature

This paper ties into multiple literatures: the one on standard form contracts, the one on personalization, the one on price discrimination, the one on the digitization of business to consumer interaction, and finally the one on the human alignment of responses given by LLMs.

Standard form contracts. Drafting a contract is a serious challenge. One must anticipate risks on the path towards contract implementation, must design and evaluate clauses meant to address these risks, assess the cost of alternative solutions, and formulate the result in a text that is likely to be upheld in court. Suppliers face a similar set of risks with most, if not all consumers. Hence for them contract design tends to exhibit considerable economies of scale. In principle, it is therefore understandable that suppliers often draft a contract, and use the same set of provisions for all customers (Griffin 1977, Patterson 2010).

Yet suppliers are not neutral arbiters. They are likely to exploit the power that contract design gives them to their individual advantage (D'Agostino 2014), which is also what consumers expect (Snyder and Mirabito 2019). Typically, the provider is not open to negotiating individual clauses (Marotta-Wurgler and Taylor 2013), giving her the power of a take-it-or-leave-it offer (Hillman and Rachlinski 2002, Patterson 2010). The content of contracts is also normally not the object of competition between suppliers (Patterson 2010). The first mover advantage is compounded by the fact that the typical consumer does not have legal training, and hence the competence to assess contract terms (D'Agostino 2014, 9), and that hardly any consumer even reads the provisions that are meant to govern her interaction with the supplier (Kessler 1943, Ayres and Schwartz 2014, Bakos, Marotta-Wurgler et al. 2014). Occasionally, suppliers even deliberately formulate contract terms that they do not mean to apply; they rather want to use them as bargaining chips in case, after the fact, the risk materializes that is covered by the clause in question (Johnston 2005). However, online fora might enable consumers to avail some counter-power, by spreading the word about unfair contract terms, and by organizing resistance (Becher and Zarsky 2007), although this too is not a very likely event (Marotta-Wurgler 2012).[12]

Personalized law. Most jurisdictions are critical about discrimination. One part of society shall not be treated less favorably, just because they share a discernible marker, like gender, race or disability. On the other hand, most jurisdictions care about giving each individual the treatment she deserves, and about effectiveness in implementing the law. With the advent of big data, the scope for sovereign intervention that targets different members of society differently has exploded. This has revived the debate over the promises and the perils of personalized law (Casey and Niblett 2019, Ben-Shahar and Porat 2021, Lemmens, Roos et al. 2024). On the one hand, personalization leads to a better match between the rule and its addressee (Ben-Shahar and Porat 2021, 43) and reduces "the production cost of precision" (Ben-Shahar and Porat 2021, 53). Rules can also be formulated differently for different addressees, such that less savvy individual addressees also stand a chance to actually understand the rule (Arbel and Becher 2022, 99ff.). A disclosure obligation can become more effective (Busch 2019), as can

12 Atamer and Pichonnaz (2020) discuss price related terms in standard form contracts, but not from the angle of consumer choice enabling the supplier to engage in price discrimination.

be default rules (Porat and Strahilevitz 2013). One might even conceive the personalization of the process of rule making (Fisher 2024).

On the other hand, personalization requires access to fine-grained individual data (Porat and Strahilevitz 2013, Ben-Shahar and Porat 2021, 212), which raises privacy issues (Acquisti, John et al. 2013). The fact that different addressees of the same rule are treated differently may be considered as discrimination (Ben-Shahar and Porat 2019, Ben-Shahar and Porat 2021, 121). There is a risk that individuals get framed, as they have once been classified, but relevant characteristics of the person or her environment have changed in the meantime (Gillis 2024). Information about individual characteristics may be used to exploit their biases (Golobardes 2022).

Price discrimination. The main concern that motivates this experiment is the potential of contract personalization to enable price discrimination. As is well understood, if a firm holding market power engages in first degree price discrimination, this is efficient. In its perfect form, price discrimination entirely removes the deadweight loss resulting from setting the monopoly price. All demand that is sufficient to cover production cost is fulfilled. Yet higher welfare comes at a distributional cost. The supplier extracts consumer rent entirely (Varian 1989). If the consumer is not perfectly rational, there is also a welfare loss (Bar-Gill 2018). Consumers widely consider price discrimination to be unfair (Kahneman, Knetsch et al. 1986, Zuiderveen Borgesius and Poort 2017).

These effects are of course perfectly standard for readers who have had some exposure to microeconomics. But in the interest of not losing anyone, let me reiterate the logic, with the help of Figure 1. In all three panels, demand is defined by the downward sloping line. The slope results from heterogeneity. Some consumers care more about the product, or have a softer budget constraint, and are willing to pay more. Others would still like to get the product or service, but only at a lower price. If producers are perfectly controlled by competition, and if at least two producers offer the identical product, they can only cover their cost, and split the revenue. If one of them sets a higher price, the other undercuts him, and gets the total revenue. This is represented by the horizontal supply curve. As some consumers would have paid more, these consumers get a rent. Actually, with perfect competition, all the gains from trade go to consumers. In the left panel, this is the large blue triangle.

If, however, there is a single producer, it can set the price at will. It will choose the price that maximizes his gains from trade. This is the red square. This distributional gain for the producer does not only reduce consumer rent: the blue triangle is much smaller, only consumers with a very high willingness to pay still receive a rent. Now there is also the grey triangle. As the price is too high for them, a fraction of consumers that could be served at the prevailing production cost do not get the product. This is why the monopoly price is inefficient.

Compare this with the right-hand panel. This panel assumes that there is not only a monopoly, but that the supplier is also perfectly informed about the individual willingness to pay of every potential buyer. This has two effects, pointing into opposite directions. The grey triangle has disappeared. Every consumer who has sufficient willingness to pay to cover production cost

is served. The inefficiency has disappeared. Yet the triangle is now completely red: all the rent goes to the producer.

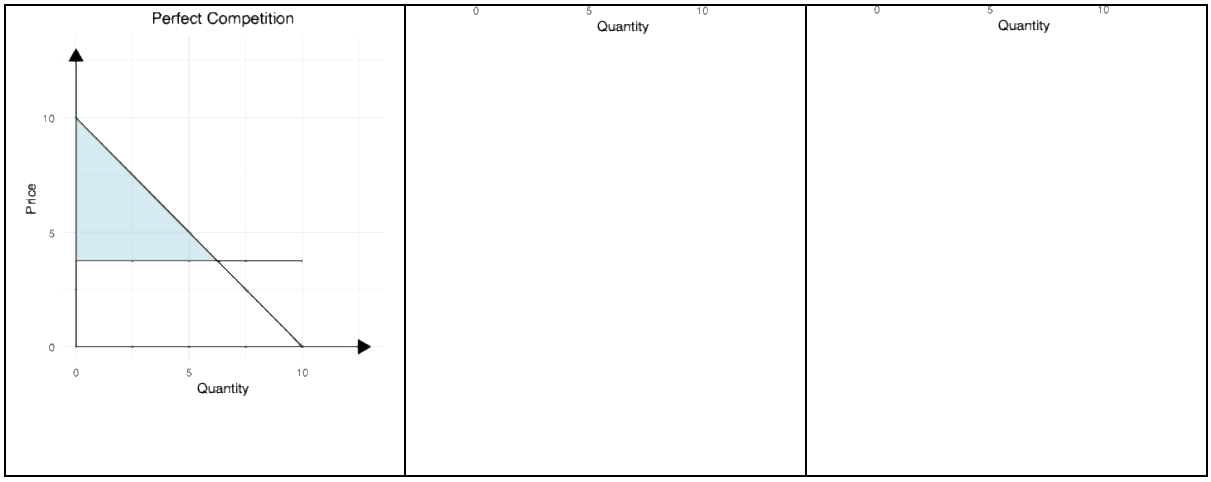

Figure 1
Logic of First Degree Price Discrimination

Human alignment of LLMs. This paper uses an LLM to estimate how customers would decide when given the choice between alternative contract clauses, and how suppliers would respond strategically. Hence the empirical strategy requires a reasonable degree of human alignment. This is why the burgeoning literature on the human alignment of LLMs is relevant for the present project.

Legal scholars have mostly been interested in the ability of LLMs to engage in legal reasoning (Guha, Nyarko et al. 2024), including taking the bar exam (Martínez 2024), and have uses LLMs for the extraction of features from legal text, to be used in quantitative analysis (Dominguez-Olmedo, Nanda et al. 2024). For this paper, literature is even more pertinent that compares the responses from LLMs to the responses of human participants on equivalent experimental designs. GPT-3 exhibits anchoring effects similar to the ones observed in humans (Jones and Steinhardt 2022), is subject to gender stereotypes (Acerbi and Stubbersfield 2023) and falls prey to intuition in cognitive reflection tests in about the same way as humans (Hagendorff, Fabi et al. 2023). GPT-3.5 exhibits moral judgements that are similar to the ones observed in human subjects (Dillion, Tandon et al. 2023) and emulates the choices well that human proposers make in the ultimatum game (Kitadai, Tsurusaki et al. 2023). However GPT-3.5 is better than human subjects at applying Bayes' rule, and is less likely to overvalue the difference between two options presented simultaneously (Orsini 2023). Finally, on multiple tasks, GPT-3.5 exhibits a "correct answer bias", such that it almost always gives the majority response, even if tested multiple times; the variance observed in human subjects on the analogous task is suppressed (Park, Schoenegger et al. 2024). GPT-4 exhibits risk preferences, time preferences and social preferences that are qualitatively similar to the ones observed in human subjects, but they are more extreme (Capraro, Di Paolo et al. 2023, Chen, Liu et al. 2023, Goli and Singh 2024).

This body of findings suggests: one should handle LLM predictions about human choices with caution. Yet as explained in the introduction, the alignment does not have to be perfect for the results to be normatively relevant. If findings are consistent across alternative contexts, and if predicted effects are sufficiently sizeable, this may well be sufficient for regulatory attention, if not regulatory intervention, especially if one considers the precautionary principle to be applicable in the relevant context.

3. Hypotheses

Hypotheses. To the best of my knowledge, offering alternatively protective contract terms as a technology for extracting consumer rent via price discrimination has not been investigated empirically. This is the topic of the present paper. The main hypothesis is:

> **H_1 contract choice:** If the supplier offers consumers a choice between alternatively protective terms of contract, it makes a higher profit.

If the supplier announces a price list for alternative degrees of protection, it must take a bet. By contrast, if the supplier waits until the customer has defined the degree of protection she requests, the supplier knows that this specific customer cares, and can set a price above the insurance cost. Moreover if the customer sees the complete pricelist, she is able to compare, and may more easily decide whether the additional degree of protection that comes with a more customer friendly clause is truly worth it. For both reasons I expect that setting the price for the clause only after the customer has chosen the degree of protection gives the firm a second mover advantage. Hence I predict:

> **H_2 second mover advantage**: The supplier makes a higher profit if it only defines the price after knowing which degree of protection the customer requests.

From the psychological literature, it is known that defaults tend to have a strong behavioral effect (Johnson and Goldstein 2003, Ben-Shahar and Pottow 2005). If this effect is critical, the fact that minimal protection is defined as the default should induce the highest number of customers to choose this version of the contract. Likewise, if maximum protection is singled out as the default, the highest fraction of customers should choose this version. If the price is higher the higher the degree of protection, with maximum protection as the default the supplier should make the highest profit. On the other hand, it is also well established that individuals evaluate losses more negatively than they evaluate gains positively; this is the main claim of prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992, Guthrie 2003, Zamir 2015). What individuals perceive as a gain, and what they perceive as a loss, depends on the way how they frame the interaction. If they define contract fulfilment as the reference point (Köszegi and Rabin 2006, Köszegi and Rabin 2007), the opposite effect would obtain: more customers demand a higher degree of protection, as otherwise they risk losing the expected benefit from concluding the contract. Hence I do not have a directed hypothesis, but test

> **H_3 default:** Supplier profit differs depending on the degree of consumer protection that is flagged out as the default.

First degree price discrimination is only possible if willingness to pay for a product or service is heterogeneous. Arguably heterogeneity is correlated with demographic markers, and with information about past choices that a consumer has made. This is why cookie data is so valuable. In a series of treatments, I ask the LLM to estimate the price the supplier is going to set if it has access to a specific feature of demographic information, and how many items the supplier expects to sell to this subgroup of population. With this data, I am able to compare the expected additional profit from such demographic information on the one hand, and information about requesting a contract that is more consumer friendly on the other hand. As the outcome of this comparison depends on the exact additional information, I have no generic hypothesis, and test these comparisons in an exploratory manner.

I do, however, have a directed hypothesis for the final set of tests. In an additional series of treatments, I combine each of these demographic markers with a contract that gives the consumer a choice between alternative contract terms. I compare each of these treatments with the companion treatment where the same demographic marker is present, but contract terms are uniform. I expect that giving the consumer a choice of contract terms has an additional effect, even if the supplier can gauge the price to the respective piece of demographic information. I therefore predict:

> **H_4 demographic information and contract choice combined:** if the supplier adds contract choice to demographic information, it makes a higher profit than when exclusively relying on the respective demographic marker.

4. Design

I have run the experiment on GPT, using version gpt-4o-2024-08-06. To see variance that makes statistical analysis meaningful, I have set “temperature” to the high value of 1. I was initially concerned that this version of the model might be too heavily tuned towards accuracy and that, therefore, I would not see sufficient variance to make statistical analysis meaningful (we have documented this issue with the original version of GPT 4 in Engel, Hermstrüwer et al. 2024). Yet if I am using GPT 3.5, the reasoning is very shallow. Happily GPT 4o does not seem to suffer from the same limitation as the original version of GPT 4, as will become visible in the results section of the paper.

For each treatment, accessing the large language model through the API, I am asking the same question 100 times.

Table 1 gives an overview over the 25 conditions. For each condition, I have 100 GPT responses for both price and quantity. Hence I have a total of 5000 independent observations.

no choice		choice	
no cookie	cookie	no cookie	cookie
base		ante	
		post	
		savings	
		upgrade	
	neighbourhood		neighbourhood_c
	flights		flights_c
	techie		techie_c
	business		business_c
	oxford		oxford_c
	pricesite		pricesite_c
	ryanair		ryanair_c
	suburb		suburb_c
	retired		retired_c
	kids		kids_c

Table 1

Treatments

I ask GPT to estimate how a firm would decide that builds on demand large scale, high end, customized screens for the presentation of electronic content. GPT learns about cost, and about two complications: occasionally, raw materials are not available. If this happens, the firm makes no profit. Moreover customization requires ad hoc adjustments of the assembly line. Production becomes considerably cheaper if the firm may sequence production freely, but may then not guarantee a delivery date.

In the *baseline*, I am refraining from also asking for quantity as otherwise I would have to give the language model detailed information about demand. That would not only make the experiment rich, with the concomitant risk of losing experimental control. I would also have to specify right from the start information about the demographics of the population of customers. That would make the treatments less clean in which I add individual pieces of demographic information. Finally, in other experiments with GPT I have found that quantitative estimates are more reliable if they are comparative. This is in line with the finding that LLMs are subject to anchoring (Jones and Steinhardt 2022). As a way out, in the *baseline* for calculating profit, I arbitrarily assume that the firm is able to sell 100 screens. In all treatments, I inform GPT that, in the absence of additional information and with a uniform price, the firm was able to sell 100 items. Hence in all treatments, there is a quantitative baseline against which GPT can gauge its estimate of quantity sold.

In the *baseline*, the firm absorbs both risks, and adjusts price. In the *choice* treatments, the firm instead gives customers the choice between three versions of the contract. In the *gold* version, the firm continues to absorb all risk. In the *silver* version, the firm does not guarantee a delivery date. In the *bronze* version, the firm reserves the right to cancel the contract.

In each of the four choice treatments, I separately ask for price and for quantity. For both questions, I give GPT a benchmark. That way I implement the functional equivalent of a within subjects design (for background see Charness, Gneezy et al. 2012). Hence for treatment comparisons I take the variance between different responses (instances of GPT) out of the equation,

and have all responses in the treatment compare their estimate with the central tendency in the *baseline* (for prices) or with an arbitrary, but constant number of sales (for quantity). Specifically, prices in the *baseline* are as follows: mean 25319, median 24400, min 20000, max 35000, sd 3343.23. In all treatments, I use the nearest prominent number, i.e. 25000, for comparison. The distribution of estimated prices is available in FigA1.

I am introducing the choice of contract clause in four different versions. In the *ante* version, the firm commits to a separate price for each of the three clauses before the customer decides. In the *post* version, the firm only specifies the price after the customer has chosen her preferred version of the contract. In two more conditions, the firm makes one of the three clauses the default. In the *upgrade* version, the bronze treatment is the default. In the *savings* version, the gold treatment is the default.

In the *cookie* conditions, I inform GPT about one piece of demographic information each. I implement five conditions that purportedly suggest a higher willingness to pay for the screen, and five conditions that suggest a lower willingness to pay. Specifically, I tell GPT that the customer lives in a wealthy neighbourhood in San Francisco; that she has taken five intercontinental business class flights last year; that she often buys the latest technology; that she runs an advertising agency; that she holds an executive MBA from Oxford. In the opposite direction, I tell GPT that this customer often goes to price checking websites; that she has last year twice booked intercontinental flights with Ryanair; that she lives in a suburb that, over the last decade, has lost a third of its population; that she is retired; that she has three kids and is the only breadwinner of the family.

In the final 10 conditions, I combine a choice of contract with each of these 10 pieces of demographic information. For the choice of contract, I am using the *post* version.

The wording of all prompts is in the Appendix.

For calculating profit, I exploit the fact that, when accessed through the API, GPT does not have memory, and when temperature is set to a high value, the portfolio of responses reflects the distribution of choices that GPT expects to find in the population. I therefore merely match estimated prices and estimated quantities by the sequence of the responses GPT has given to the respective prompt. One may object that, arguably, the estimates of price and quantity would be correlated, were I to elicit them simultaneously. I have not done so as otherwise for GPT the task would have been considerably more involved. I would actually have expected GPT to engage in business planning, trading more sales against lower prices, and vice versa. As I wanted to introduce a choice of contract clauses, this would have meant a three-dimensional choice. The latest version of GPT (GPT 4o1) prides itself of having more advanced reasoning capabilities. But in my project I am not per se interested in the accurate handling of a complex problem. What I want to learn is the likely distribution of choices in the population. This is why I have preferred to separately ask for price and quantity, and to only match results thereafter. I acknowledge the possibility that some of these matches are less plausible. But this limitation is held constant and should not affect the estimation of treatment effects, which is what I want to study.

5. Results

a) Effect of Contract Choice

Figure 2 has the main result: if the firm gives consumers a choice between the three different contracts, this has a clear positive effect on its profit.

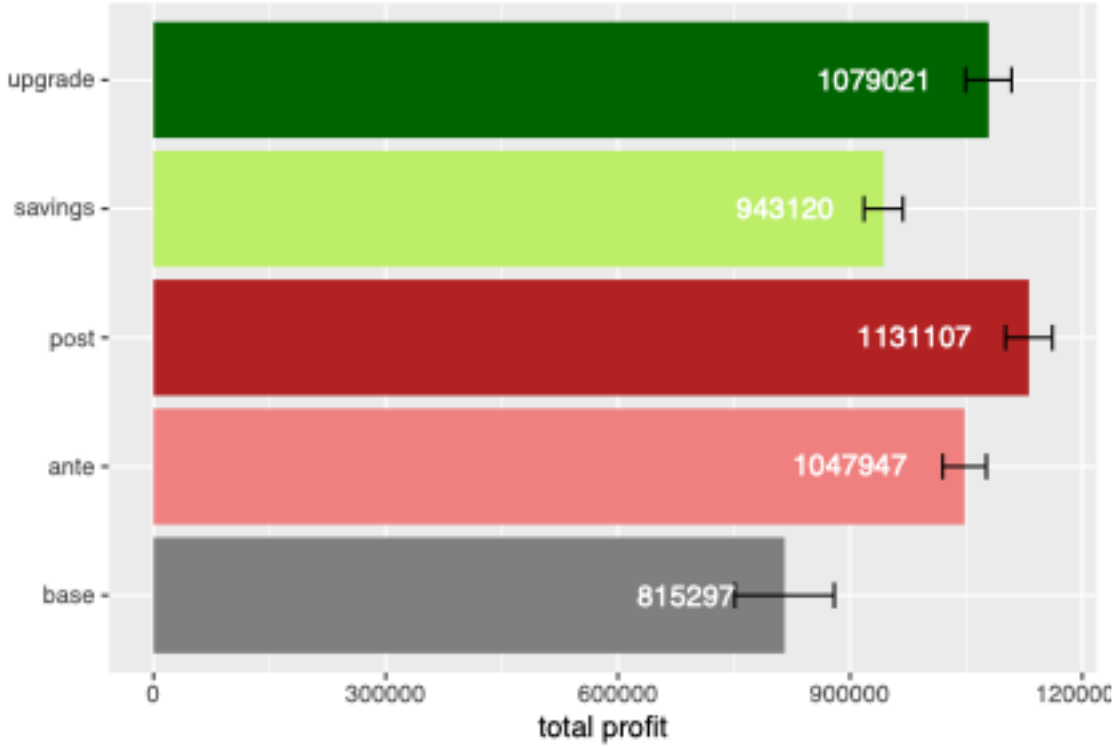

Figure 2
Effect of Contract Choice on Profit

The regression in Table 2 shows that, irrespective of treatment, profit is significantly higher if the supplier gives the customer a choice between three different contract terms.

ante	232649*** (27673)
post	315809*** (27673)
savings	127823*** (27673)
upgrade	263724*** (27673)
cons	815297*** (19568)
N	500

Table 2
Effect of Contract Choice on Profit
OLS
reference category: baseline
standard errors in parenthesis
*** p < .001

Consequently, hypothesis $\mathbf{H_1}$ is clearly supported by the data. I conclude

> **Result_1: contract choice:** If the supplier offers consumers a choice between alternatively protective terms of contract, it makes a higher profit.

With the help of Wald tests of the regression in Table 2, I can also test $\mathbf{H_2}$: the profit in the *post* condition is significantly higher than the profit in the *ante* condition, $p < .001$. This gives me

> **Result_2: second mover advantage**: The supplier makes a higher profit if it only defines the price after knowing which degree of protection the customer requests.

Another Wald test on the same regression shows that the profit in the *upgrade* condition is significantly higher than in the *savings* condition, $p < .001$. Hence (overall) I do not find a default effect. The result is consistent with aversion against the risk of losing the benefit from contract fulfilment. This gives me

> **Result_3: aversion against losing the benefit from contract fulfilment**: If full protection is flagged out as the default, the producer makes a lower profit than if minimal protection is the default.

As Figure 3 shows, in the *savings* condition, suppliers make a lot of money with selling the gold version, while in the *upgrade* condition, they earn most with the bronze version of the contract. This suggests a clear default effect. The main reason why, overall, suppliers do worse in the *savings* condition are earnings from the bronze contract: they are tiny, compared with all other types of contract. Table A1 shows that two effects compound: firms sell few bronze contracts, and they sell them at a much lower price than the same contract sells in the *upgrade* condition.

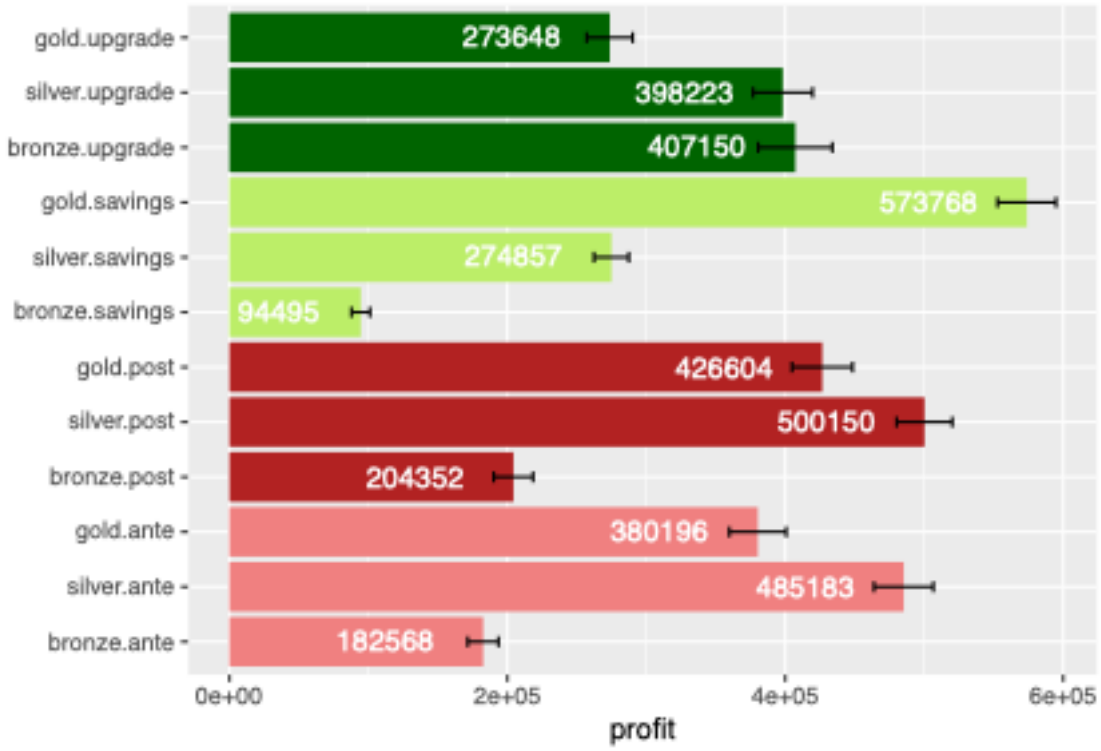

Figure 3
Effect of Contract Choice on Profit, By Chosen Contract

The difference in earnings between the *ante* and the *post* condition is more subtle (see TableA1). If the firm waits with defining the price until the customer has chosen her preferred version of the contract, prices for the silver and gold version are a bit higher. The combination of both effects drives the second-mover advantage.

b) Demographic Information

Figure 4 shows that demographic information has two clear effects: it substantially reduces the uncertainty about the price a firm should set (error bars are much smaller with demographic information), and it tells the firm whether it could expect to sell at a markup or whether, by contrast, it must reduce price if it wants to maximize profit. All estimates have the expected sign: cookie information likely positively correlated with higher income (neighbourhood, flights, Oxford) or greater interest in the latest technology (techie, business) makes the firm expect that it can sell at a higher price, and make a higher profit. By contrast if the customer is retired, must fend for several kids, lives in a poorer neighbourhood, she likely has a lower income. And if she has repeatedly used Ryanair for long-distance flights, or often goes to price checking sites, this customer is likely more price sensitive than the average customer.[13]

13 As I have separately elicited estimates for price and for quantity, there are technically two alternative versions for calculating profit: multiplying the estimated price with selling 100 items; selling the estimated number of screens at the average price when no demographic information about the individual customer is available (25000). In Figure 4 I am using the version based on estimated sales. In Figure A2 I also report the alternative calculations. Table A2 provides statistical evidence.

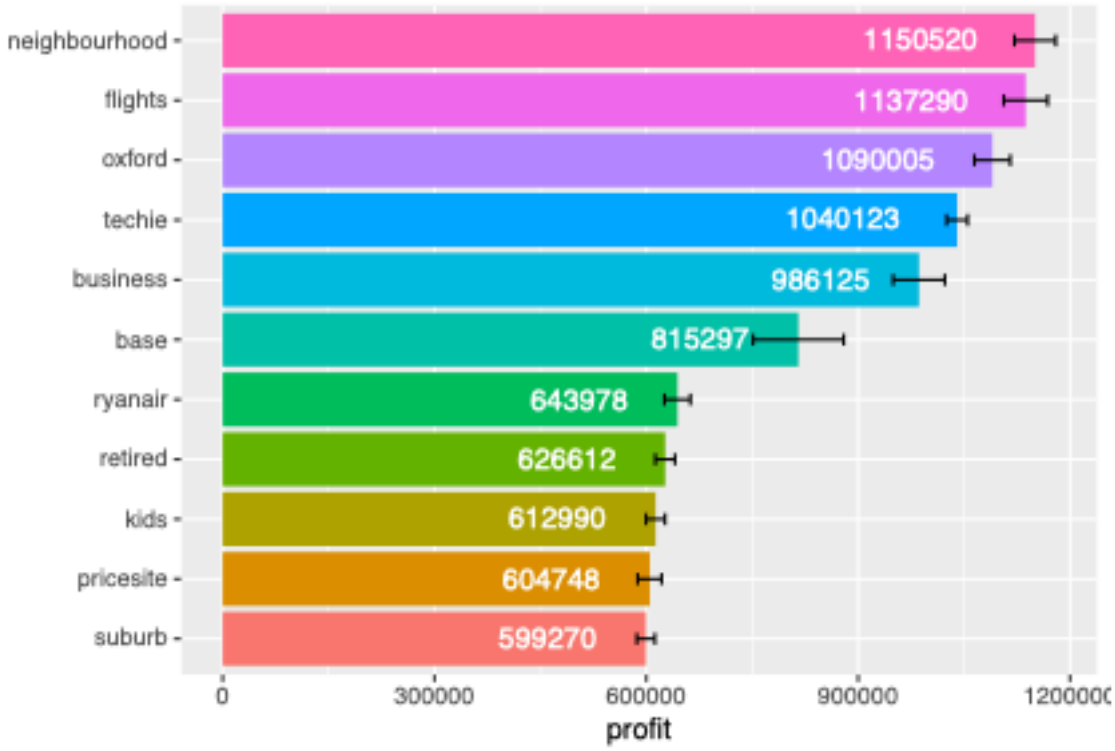

Figure 4

Profit in the Absence on Contract Choice

c) Choice vs. Demographic Information

Figure 5 is even more eye opening: the mere fact of giving customers a choice is almost as effective as the most informative demographic information, and not significantly different from these pieces of information[14]. Hence even if the firm does not have access to cookie information, it can achieve a comparable increase in profit by simply giving the customer a choice between different contract terms.

14 See the regression in Table A3.

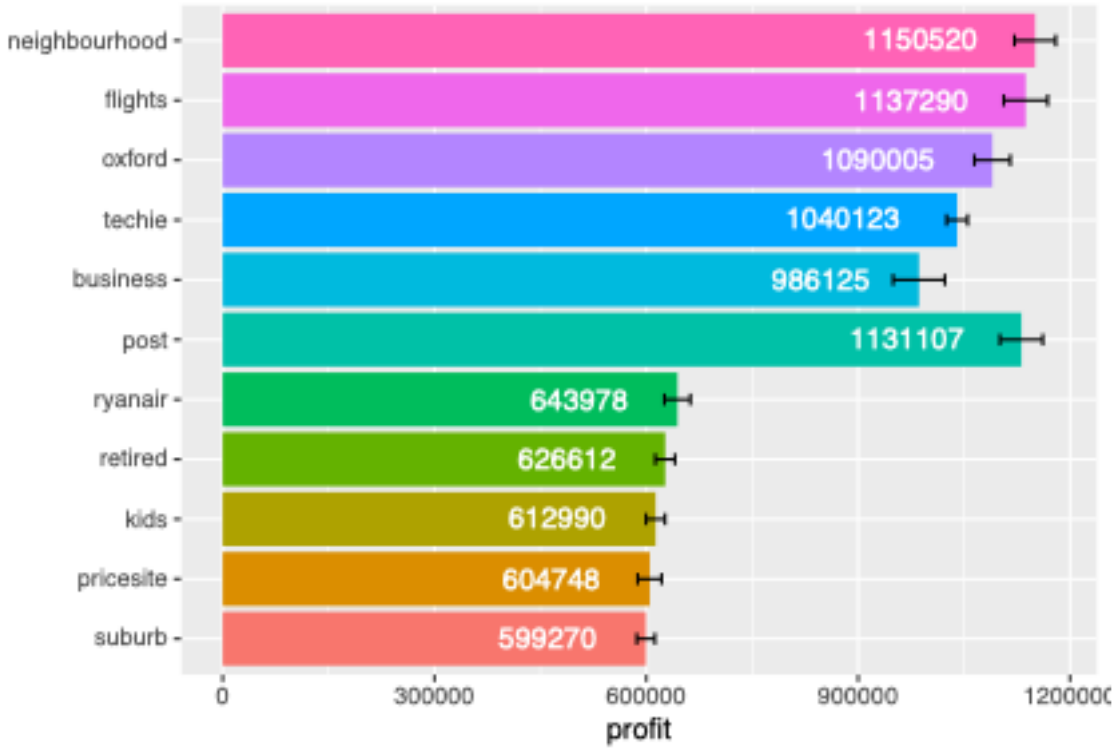

Figure 5
Choice vs. Demographic Information: Effect on Profit

d) Demographic Information and Choice

In the final step of data analysis, I compare the effect of each piece of demographic information per se on profit with the combined effect of this piece of cookie data with giving the customer a choice between contract terms. As Figure 6 shows, for each and every different piece of demographic information, this has an effect: the combination always increases profit. Statistics are in Table 3. For *pricesite*, which I have chosen as the reference category, the effect of adding contract choice is very strong (profit is 30% higher). As the interaction effect shows, it is only even bigger in the *neighbourhood* condition. For all other pieces of demographic information, the interaction effect is negative, indicating that the increase in profit resulting from giving the consumer a choice is smaller than in the *pricesite* condition. Yet subsequent Wald tests show that the difference is always significantly different from zero. Whatever demographic information is available to the seller, additionally giving the consumer a choice of contract terms is always profitable.

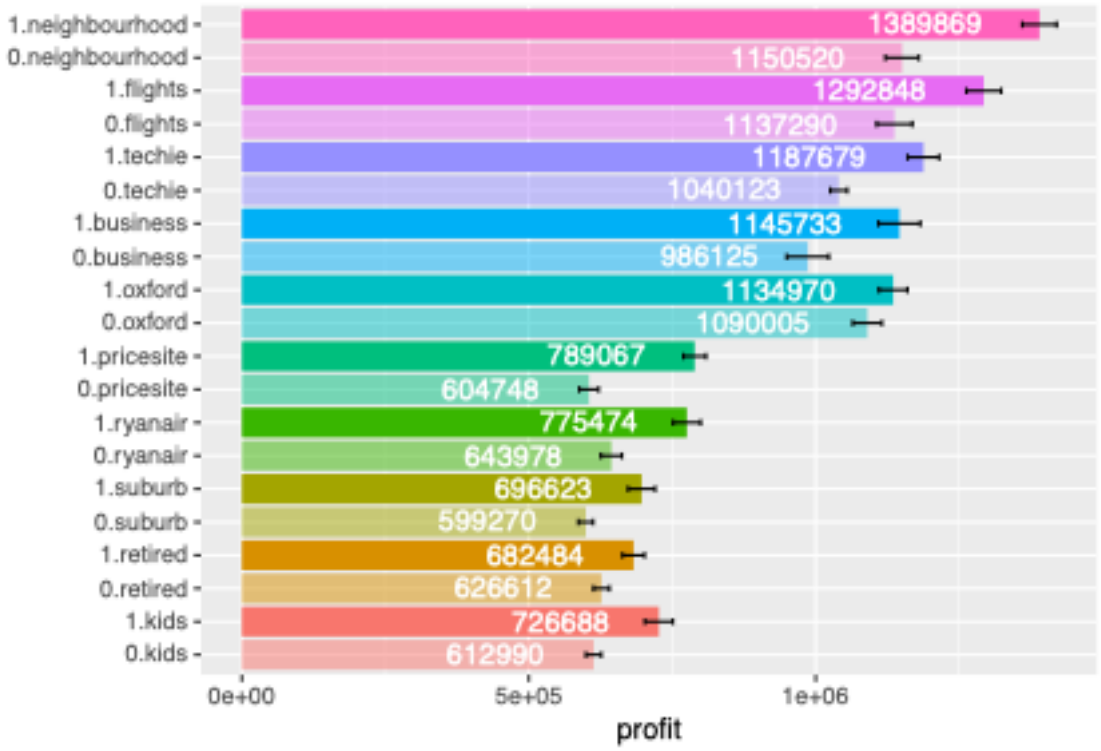

Figure 6

Combination of Choice with Demographic Information

	profit
neighbourhood	545772*** (17831)
flights	532542*** (17831)
techie	435375*** (17831)
business	381377*** (17831)
oxford	485257*** (17831)
ryanair	39229* (17831)
suburb	-5478 (17831)
kids	8242 (17831)
retired	21864 (17831)
choice	184319*** (17831)
neighbourhood*choice	55030* (25217)
flights*choice	-28761 (17831)
techie*choice	-36763 (17831)
business*choice	-24711 (17831)
oxford*choice	-139354*** (17831)
ryanair*choice	-52823* (17831)
suburb*choice	-86966*** (17831)
kids*choice	-70621** (17831)
retired*choice	-128447*** (17831)
cons	604748*** (12609)
N	2000

Table 3

Combined Effect of Demographic Information and Contract Choice

OLS

reference category: pricesite

standard errors in parenthesis

*** p < .001, ** p < .01, * p < .05

I thus also support hypothesis **H_4** and note

> **$Result_4$: demographic information and contract choice combined:** if the supplier adds contract choice to demographic information, it makes a higher profit than when exclusively relying on the respective demographic marker.

6. Discussion

Summary. In an experiment on the large language model GPT-4o, a supplier always makes a higher profit if it replaces uniform contract terms with a set of terms between which the customer may choose. The extra profit results from price discrimination. There is a first order and a second order effect. The first order effect results from heterogeneous willingness to pay for a more protective term. The second order effect results from the possibility that contract choice is a signal for general willingness to pay for the traded commodity. In the experiment, the effect is bigger if the least protective version is labelled as the default, and more protective terms as an "upgrade". The effect is smaller if, conversely, the most protective version is labelled as the default and less protective (and cheaper) versions as an opportunity for "savings". The effect is also bigger if the supplier only sets the price after it knows which version of the contract the consumer chooses. Hence there is a clear second mover advantage.

The profit increasing effect of giving the consumer a choice is strong. There is no piece of demographic information that has a stronger effect. Most pieces of demographic information (which the supplier might, for instance, learn from cookie data) have a significantly smaller effect on profit. If the supplier combines cookie information about demographic markers with contract choice, it always makes an extra profit. Hence even if the supplier already has access to information that allows for price discrimination, there is an additional effect from providing contract choice. Practically this is particularly appealing for the supplier if the demographic information suggests that general willingness to pay is rather low.

Testing an ideal case. Every experiment has limitations, and this experiment is no exception. In a way, the experiment tests an ideal case. The commodity in question is a screen that is customized to the specific wishes of the individual consumer. The provider does not face serious competition. This not only makes price discrimination appealing for the supplier. It also makes it difficult for customers to protect themselves. Arbitrage is hindered by customization. As arguably each individual screen is different, it is also not easy for consumers to learn how other consumers have been treated. Given production is on demand, it is not obviously unfair that the provider tries to reduce cost by additional degrees of freedom during contract implementation. If one is concerned that effects might be less pronounced under less ideal circumstances, one would have to test additional cases. With the help of a large language model happily this is a feasible prospect.

Which contract clauses are most informative? A typical standard form contract addresses a host of ways in which implementation can go wrong.[15] As consumers next to never read terms and conditions (Bakos, Marotta-Wurgler et al. 2014), they are also unlikely to wade through long lists of alternative clauses. Moreover different alternative versions of more than one clause might have offsetting or compounding effects on producer profit. For both reasons, producers can practically only give consumers very limited choice. In a follow-up, I have asked

15 For an illustration, consider the standard form contract Apple uses to sell its products in Europe, https://www.apple.com/legal/procurement/docs/OL-APAC-AP_v.1.0.pdf,

GPT-4o consumer choice to rank the following five clauses in terms of their diagnostic value for willingness to pay:[16]

a) cancellation of contract for exceptional increase in cost ruled out

b) guarantee of delivery date

c) test of functionality upon delivery

d) warranty period extended to 2 years

e) no use of purchase data in future negotiations

I have used the same LLM as in the main experiment, through the API, have set temperature to 1 to see the expected variance, and have asked 100 times. As Figure 7 shows, GPT has clear opinions. It expects the two clauses that I have tested in my experiment (cancellation, [equivalent to *bronze*] and leeway with the delivery date [equivalent to *silver*]) to be most diagnostic, more encompassing data protection to be least diagnostic, and a functionality test upon deliver, or an extension of the warranty period, to be in between.

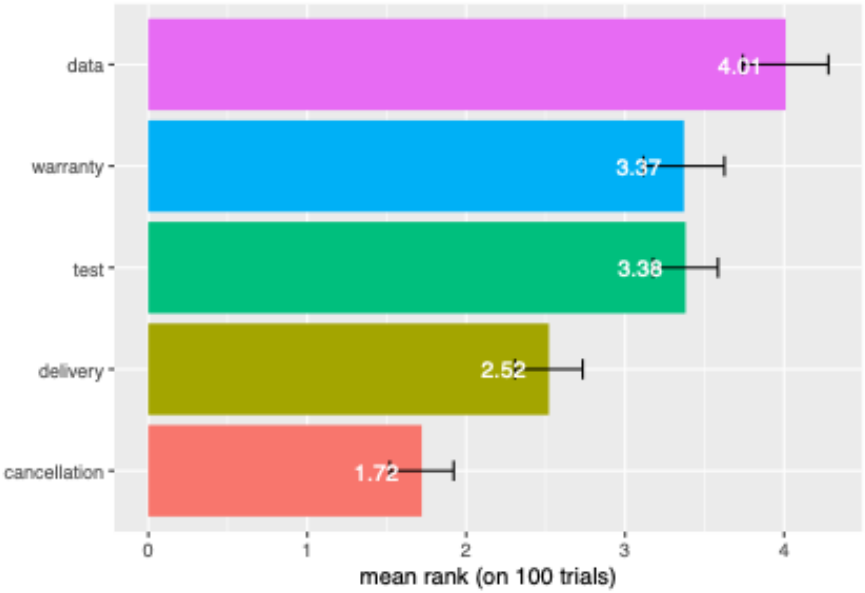

Figure 7
Diagnostic Value of Alternative Choices of Contract Clauses

Human alignment. Large language models have been trained on more human utterances than any living human being could ever hear or read in their lifetime. Large language models can also be programmed to produce an entire distribution of choices. This not only makes statistical analysis meaningful. It also informs normative debates about the scope of a policy problem, and the way how it is distributed in the population of interest. Finally experiments on large language models can be automated, are rapid and affordable. This is why they open the door

16 For the complete prompt, see Appendix ***.

towards generating much richer evidence than could ever be collected with human participants.

These advantages notwithstanding, large language models are of course not human beings. Whether the results received from these models are a reliable proxy of human behavior is still very much an open question. While in some contexts human alignment seems to be reasonably good, in other contexts responses deviate from what is known about typical human choices. One should therefore treat any experiment on large language models, and certainly the present experiment, with caution.

But in this specific case, the biggest advantage is observability. With real customers, at best one might convince a supplier to offer two different contracts to two randomly selected subgroups of consumers. Such a randomly controlled trial would certainly be interesting, in particular due to its external validity. But strictly speaking if the same supplier offers different contracts to different consumers, by the fact that the identity of the supplier is held constant, there is dependence. More importantly, even in such a best scenario, one could perhaps study one single variation to the uniform contract. By contrast, with the help of a large language model, one is able to test a rich set of parameter combinations. One is therefore in a position to generate much more robust evidence, and to find out which specific features have the biggest effect. This advantage is particularly valuable for regulators, and for customers or customer organizations planning to sue suppliers. While suppliers can relatively easily, and confidentially, run A/B tests, actual A/B tests are close to impossible for regulators. By contrast, as this papers shows, LLM experiments simulating such tests are feasible and informative.

Normative conclusion. In the experiment, offering a choice of contract terms gives the supplier a higher profit. This profit results from first degree price discrimination. As always with price discrimination by a monopolist, the demand of more consumers is served. Hence there is an efficiency gain. However it comes at a distributional disadvantage for consumers. Now this disadvantage results from contract design. Does the power asymmetry resulting from the fact that one supplier interacts with a population of consumers justify legal intervention? Standard form contract law traditionally deals with a different normative concern. It deals with exploitation resulting from the imposition of an unfavourable term. The results of the present experiment suggest that there is a gap. The extractive effect of contract choice also deserves normative legal attention. The results from the present experiment suggest that the law should prevent consumers from falling into the negotiation trap.

***

Differences in the level of protection notwithstanding, all jurisdictions agree that special protection is required if the supplier has preformulated the terms of contract.[17] Within the general limits of private law, consumers remain free to negotiate a contract that deviates from the default rules of private law, and that would not be legal in a standard form contract. The logic of the exception is straightforward: as the parties have individually negotiated the clause in

17 § 305b BGB; Art. 3 (1) Directive 93/13/EC; for US law see only *Thompson Crane & Trucking Co. v. Eyman, 267 P.2d 1043 (Cal. Dist. Ct. App. 1954).*

question, it can be assumed that voluntary consent is credible. Again jurisdictions have been differently protective in the interpretation of the term negotiation. Under German law, the negotiation exception in § 305 I 3 BGB does not apply if a seller offers alternative versions of a contract clause from which the buyer may choose. As long as the content has been preformulated, the clause is still regarded as part of a standard form contract (BGH NJW-RR 2018, 814).

References

Acerbi, Alberto and Joseph M Stubbersfield (2023). "Large Language Models Show Human-Like Content Biases in Transmission Chain Experiments." *Proceedings of the National Academy of Sciences* **120**(44): e2313790120.

Acquisti, Alessandro, Leslie K John and George Loewenstein (2013). "What Is Privacy Worth?" *Journal of Legal Studies* **42**(2): 249-274.

Aguinis, Herman and Kyle J Bradley (2014). "Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies." *Organizational Research Methods* **17**(4): 351-371.

Angrist, Joshua D and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics. An Empiricist's Companion*. Princeton, Princeton University Press.

Arbel, Yonathan A and Shmuel I Becher (2022). "Contracts in the Age of Smart Readers." *George Washington Law Review* **90**: 83-146.

Atamer, Yeşim M and Pascal Pichonnaz (2020). *Control of Price Related Terms in Standard Form Contracts*, Springer.

Atzmüller, Christiane and Peter M Steiner (2010). "Experimental Vignette Studies in Survey Research." *Methodology*.

Ayres, Ian and Alan Schwartz (2014). "The No-Reading Problem in Consumer Contract Law." *Stanford Law Review* **66**: 545-610.

Bakos, Yannis, Florencia Marotta-Wurgler and David R Trossen (2014). "Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts." *Journal of Legal Studies* **43**(1): 1-35.

Bar-Gill, Oren (2018). "Algorithmic Price Discrimination When Demand Is a Function of Both Preferences and (Mis)Perceptions." *University of Chicago Law Review* **86**: 217-254.

Becher, Shmuel I and Tal Z Zarsky (2007). "E-Contract Doctrine 2.0. Standard Form Contracting in the Age of Online User Participation." *Michigan Telecommunications and Technology Law Review* **14**: 303-366.

Ben-Shahar, Omri and Ariel Porat (2019). "Personalizing Mandatory Rules in Contract Law." *University of Chicago Law Review* **86**: 255-282.

Ben-Shahar, Omri and Ariel Porat (2021). *Personalized Law. Different Rules for Different People*. New York, Oxford University Press.

Ben-Shahar, Omri and John AE Pottow (2005). "On the Stickiness of Default Rules." *Florida State University Law Review* **33**: 651-682.

Busch, Christoph (2019). "Implementing Personalized Law." *University of Chicago Law Review* **86**(2): 309-332.

Capraro, Valerio, Roberto Di Paolo and Veronica Pizziol (2023). "Predict-Ai-Bility of How Humans Balance Self-Interest with the Interest of Others."

Casey, Anthony J and Anthony Niblett (2019). "Framework for the New Personalization of Law." *University of Chicago Law Review* **86**(2): 333-358.

Charness, Gary, Uri Gneezy and Michael A Kuhn (2012). "Experimental Methods: Between-Subject and within-Subject Design." *Journal of economic behavior & organization* **81**(1): 1-8.

Chen, Yiting, Tracy Xiao Liu, You Shan and Songfa Zhong (2023). "The Emergence of Economic Rationality of Gpt." *Proceedings of the National Academy of Sciences* **120**(51): e2316205120.

D'Agostino, Elena (2014). *Contracts of Adhesion between Law and Economics. Rethinking the Unconscionability Doctrine*, Springer.

Dillion, Danica, Niket Tandon, Yuling Gu and Kurt Gray (2023). "Can Ai Language Models Replace Human Participants?" *Trends in Cognitive Sciences* **27**(7): 597-600.

Dominguez-Olmedo, Ricardo, Vedant Nanda, Rediet Abebe, Stefan Bechtold, Christoph Engel, Jens Frankenreiter, Krishna P Gummadi, Moritz Hardt and Michael A Livermore (2024). Lawma: The Power of Specialization for Legal Tasks.

Engel, Christoph, Yoan Hermstrüwer and Alison Kim (2024). Do Algorithmic Decision-Aids Disempower Democracy and the Rule of Law? An Experiment with Large Language Models.

Engel, Christoph and Richard H. McAdams (2024). "Asking Gpt for the Ordinary Meaning of Statutory Terms." *Journal of Law, Technology & Policy*: 235-296.

Fisher, Talia (2024). "Personalizing Personalized Law: Discussion of “Personalized Law”." *Jerusalem Review of Legal Studies* **29**(1): 32-47.

Foster, Kenneth R, Paolo Vecchia and Michael H Repacholi (2000). "Science and the Precautionary Principle." *Science* **288**(5468): 979-981.

Gillis, Talia B (2024). "Unstable Personalized Law." *Jerusalem Review of Legal Studies* **29**(1): 65-84.

Goli, Ali and Amandeep Singh (2024). "Frontiers: Can Large Language Models Capture Human Preferences?" *Marketing Science*.

Golobardes, Mireia Artigot (2022). "Algorithmic Personalization of Consumer Transactions and the Limits of Contract Law." *Journal of Law, Market & Innovation* **1**(1): 17-43.

Griffin, Ronald C (1977). "Standard Form Contracts." *North Carolina Central Law Journal* **9**: 158-177.

Guha, Neel, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore and Diego Zambrano (2024). "Legalbench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* **36**.

Guthrie, Chris (2003). "Prospect Theory, Risk Preference, and the Law." *Northwestern University Law Review* **97**: 1115-1163.

Hagendorff, Thilo, Sarah Fabi and Michal Kosinski (2023). "Human-Like Intuitive Behavior and Reasoning Biases Emerged in Large Language Models but Disappeared in Chatgpt." *Nature Computational Science* **3**(10): 833-838.

He, Yu-Lin, Xiao-Liang Zhang, Wei Ao and Joshua Zhexue Huang (2018). "Determining the Optimal Temperature Parameter for Softmax Function in Reinforcement Learning." *Applied Soft Computing* **70**: 80-85.

Hillman, Robert A and Jeffrey J Rachlinski (2002). "Standard-Form Contracting in the Electronic Age." *New York University Law Review* **77**: 429-495.

Johnson, Eric J. and Daniel Goldstein (2003). "Do Defaults Save Lives?" *Science* **302**: 1338-1339.

Johnston, Jason Scott (2005). "The Return of Bargain. An Economic Theory of How Standard-Form Contracts Enable Cooperative Negotiation between Businesses and Consumers." *Michigan Law Review* **104**: 857-898.

Jones, Erik and Jacob Steinhardt (2022). "Capturing Failures of Large Language Models Via Human Cognitive Biases." *Advances in Neural Information Processing Systems* **35**: 11785-11799.

Kahneman, Daniel, Jack L Knetsch and Richard Thaler (1986). "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review* **76**(4): 728-741.

Kahneman, Daniel and Amos Tversky (1979). "Prospect Theory. An Analysis of Decision under Risk." *Econometrica* **47**: 263-291.

Kessler, Friedrich (1943). "Contracts of Adhesion--Some Thoughts About Freedom of Contract." *Columbia Law Review* **43**(5): 629-642.

Kitadai, Ayato, Yudai Tsurusaki, Yusuke Fukasawa and Nariaki Nishino (2023). *Toward a Novel Methodology in Economic Experiments: Simulation of the Ultimatum Game with Large Language Models*. 2023 IEEE International Conference on Big Data (BigData), IEEE.

Kornhauser, Lewis A (1976). "Unconscionability in Standard Forms." *California Law Review* **64**: 1151-1183.

Korobkin, Russell (2003). "Bounded Rationality, Standard Form Contracts, and Unconscionability." *University of Chicago Law Review* **70**: 1203-1296.

Köszegi, Botond and Matthew Rabin (2006). "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* **121**: 1133-1165.

Köszegi, Botond and Matthew Rabin (2007). "Reference-Dependent Risk Attitudes." *American Economic Review* **97**(4): 1047-1073.

Kriebel, David, Joel Tickner, Paul Epstein, John Lemons, Richard Levins, Edward L Loechler, Margaret Quinn, Ruthann Rudel, Ted Schettler and Michael Stoto (2001). "The Precautionary Principle in Environmental Science." *Environmental health perspectives* **109**(9): 871-876.

Larsen, Nicholas, Jonathan Stallrich, Srijan Sengupta, Alex Deng, Ron Kohavi and Nathaniel T Stevens (2024). "Statistical Challenges in Online Controlled Experiments: A Review of a/B Testing Methodology." *American Statistician* **78**(2): 135-149.

Leff, Arthur Allen (1967). "Unconscionability and the Code—the Emperor's New Clause." *Pennsylvania State Law Review* **115**: 485-559.

Lemmens, Aurelie, Jason MT Roos, Sebastian Gabel, Eva Ascarza, Hernan Bruno, Brett R Gordon, Ayelet Israeli, Elea McDonnell Feit, Carl F Mela and Oded Netzer (2024). Personalization and Targeting: How to Experiment, Learn & Optimize.

Llewellyn, Karl N (1960). *The Common Law Tradition: Deciding Appeals*. Boston, Little Brown.

Marotta-Wurgler, Florencia (2012). "Does Contract Disclosure Matter?" *Journal of Institutional and Theoretical Economics* **168**: 94-119.

Marotta-Wurgler, Florencia and Robert Taylor (2013). "Set in Stone. Change and Innovation in Consumer Standard-Form Contracts." *New York University Law Review* **88**: 240-285.

Martínez, Eric (2024). "Re-Evaluating Gpt-4's Bar Exam Performance." *Artificial Intelligence and Law*: 1-24.

Orsini, Elia (2023). *Do Cognitive Biases Persist in Large Language Models?*

Park, Peter S, Philipp Schoenegger and Chongyang Zhu (2024). "Diminished Diversity-of-Thought in a Standard Large Language Model." *Behavior Research Methods*: 1-17.

Patterson, Mark R (2010). "Standardization of Standard-Form Contracts: Competition and Contract Implications." *William and Mary Law Review* **52**: 327-414.

Porat, Ariel and Lior Jacob Strahilevitz (2013). "Personalizing Default Rules and Disclosure with Big Data." *Michigan Law Review* **112**: 1417-1478.

Snyder, Franklin G and Ann M Mirabito (2019). "Boilerplate: What Consumers Actually Think About It." *Ind. L. Rev.* **52**: 431-454.

Sunstein, Cass R (2015). *Choosing Not to Choose: Understanding the Value of Choice*, Oxford University Press, USA.

Thomas, Jeffrey E (1998). "An Interdisciplinary Critique of the Reasonable Expectations Doctrine." *Connecticut Insurance Law Journal* **5**: 295-334.

Tversky, Amos and Daniel Kahneman (1992). "Advances in Prospect Theory. Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* **5**: 297-323.

Varian, Hal R. (1989). "Price Discrimination." *Handbook of industrial organization* **1**: 597-654.

Zamir, Eyal (2015). *Law, Psychology, and Morality. The Role of Loss Aversion*, Oxford University Press, USA.

Zamir, Eyal and Ian Ayres (2020). "A Theory of Mandatory Rules: Typology, Policy, and Design." *Texas Law Review* **99**: 283-340.

Zuiderveen Borgesius, Frederik and Joost Poort (2017). "Online Price Discrimination and Eu Data Privacy Law." *Journal of consumer policy* **40**: 347-366.

Appendix

I. Main Experiment: Prompts

I am holding the system prompt constant throughout the entire experiment, as follows:

System prompt:

"##### Issue

A firm produces large video screens to the exact specifications defined by the client: size; technology (LCD, LED; OLED); brightness; number of pixels per square centimeter; repeat rate; energy consumption; longevity.

The firm sells the screens through a dedicated website. On the website there is a link to "terms and conditions", which cover all the typical elements in a production contract.

Payment is due upon delivery.

The cost of producing the screen ordered by customer C is 20,000 €. The firm has no serious competitor.

For producing the screens,
a) the firm needs raw materials that in the past sometimes had been hard to get. Alternative raw materials cause a clear loss in quality. If this happens, the firm would have to renegotiate the contract. If the customer is not willing to accept lower quality, the firm would have to compensate the customer for the loss she has incurred. The firm estimates the probability of such a shortage to be 2%.
b) the production facility must be adjusted to every new screen, and depending on the remaining contracts, some specifications may be harder to squeeze in. Based on its past experiences, the firm estimates that, with reasonable flexibility in sequencing separate offers, the production cost is 20% lower.

Task

Please fulfil the task defined in the user prompt.

Reasoning process

Please first explain in natural language

a) in which criteria the firm should ground its decision
b) how important each of these criteria is for the decision

Format
Please then respond with a number, in JSON format, using

{
"suggested_price": <number>
}

where, of course, <number> is to be replaced with your estimate."

User prompt in baseline. In the baseline, I am only asking for price. The prompt reads:

"How much should the firm charge if it has no information about the customer?"

Choice treatments. In the choice treatments, user prompts for prices read:

price, all treatments: "The firm responds to these circumstances by an alternative offer:
a) "gold": delivery is guaranteed, with all specifications, at a defined date
b) "silver": delivery is guaranteed, with all specifications, but the firm reserves the right to adjust the originally indicated delivery date
c) "bronze": the firm reserves the right to cancel the contract

ante: The website is programmed such that the firm defines its price before the customer has fully specified the order, including the contract provisions. Hence the firm commits to its price before communicating with the customer.

post: The website is programmed such that the firm only defines its price once the customer has fully specified the order, including the contract provisions.

upgrade: The website is programmed the following way:
a) the "silver" and "gold" options are flagged out as "upgrades". "Bronze" is preticked. The customer must actively change it when choosing "silver" or "gold".
b) the firm only defines its price once the customer has fully specified the order, including the contract provisions.

savings: The website is programmed the following way:
a) the "silver" and "bronze" options are flagged out as "savings options". "Gold" is preticked. The customer must actively change it when choosing "silver" or "bronze".
b) the firm only defines its price once the customer has fully specified the order, including the contract provisions.

all treatments: Assume that before the firm had introduced these options, and if it knew nothing about the customer, it had charged 25000 for a screen with the technical specifications the customer wishes to buy. Which prices should it charge if the customer, in addition, has chosen the "gold", "silver" or "bronze" option?

Please use the following JSON format for the response:

{
"suggested_price_gold": <number>,
"suggested_price_silver": <number>,
"suggested_price_bronze": <number>
}"

For quantity, the prompts read:

quantity, all treatments: “The firm responds to these circumstances by an alternative offer:
a) "gold": delivery is guaranteed, with all specifications, at a defined date
b) "silver": delivery is guaranteed, with all specifications, but the firm reserves the right to adjust the originally indicated delivery date
c) "bronze": the firm reserves the right to cancel the contract

<treatments defined the same way as for price>

Assume that before the firm had introduced these options, and if it knew nothing about the customer, it had charged 25000 for a screen with the technical specifications the customer wishes to buy, and had been able to sell 100 screens with comparable specifications.

What do you think: how will the availability of the three contract options affect the total number of screens sold? Will this total number stay the same, decrease or increase?

In the next step, please estimate how many of the new total number of sales will come with each of the three contract options? Please make sure that your estimates for "gold", "silver" and "bronze" add up to your estimate for "total".

Please use the following JSON format for the response:

```
{
  "estimated_sales_total": <number>,
  "estimated_sales_gold": <number>,
  "estimated_sales_silver": <number>,
  "estimated_sales_bronze": <number>
}”
```

Cookie data treatments with uniform contract. In the first set of cookie data treatments, in every treatment I give GPT one additional piece of information, purportedly gleaned from access to cookie data.

price, all treatments. “Assume the firm charges 25000 if it knows nothing about the customer. How much should it charge if it knows from cookie data that the concrete customer

neighbourhood: lives in a wealthy neighbourhood in San Francisco?

flights: has last year booked five intercontinental business class trips?

techie: often buys the latest technology very shortly after it has been released?

business: is an advertising agency?

oxford: holds an Oxford Executive MBA?

pricesite: goes to price checking websites?

ryanair: has last year booked two long-distance flights with Ryanair?

suburb: lives in a suburb that has lost a third of its population over the last decade?

retired: is retired?

kids: has three kids, and is the only breadwinner in the family?"

For quantity, I am informing GPT

quantity, all treatments: "Assume that as long as the firm knew nothing about the customer, it had charged 25000 for a screen with the technical specifications the customer wishes to buy, and had been able to sell 100 screens with comparable specifications.

What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who

<treatments defined the same way as for price>

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup.

Please use the following JSON format for the response:

{
"estimated_sales_total": <number>,
}"

Cookie data treatments with contract choice. In the cookie data treatments, all manipulations are contrasted with the "post" version of choice. The user prompt for the elicitation of prices reads:

price, all treatments. "The firm responds to these circumstances by an alternative offer:
a) "gold": delivery is guaranteed, with all specifications, at a defined date
b) "silver": delivery is guaranteed, with all specifications, but the firm reserves the right to adjust the originally indicated delivery date
c) "bronze": the firm reserves the right to cancel the contract

The website is programmed such that the firm only defines its price once the customer has fully specified the order, including the contract provisions.

Assume that before the firm had introduced these options, and if it knew nothing about the customer, it had charged 25000 for a screen with the technical specifications the customer wishes to buy. Which prices should it charge if

a) the firm knows from cookie data that the concrete customer

<treatments defined the same way as with uniform price>

b) the customer has chosen the "gold", "silver" or "bronze" option

Please use the following JSON format for the response:

{
"suggested_price_gold": <number>,
"suggested_price_silver": <number>,
"suggested_price_bronze": <number>
}

For quantity, the prompts read:

quantity, all treatments: Assume that as long as the firm knew nothing about the customer, it had charged 25000 for a screen with the technical specifications the customer wishes to buy, and had been able to sell 100 screens with comparable specifications.

sales to a subgroup

neighbourhood: In a first step, I had asked 100 times: "What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who live in a wealthy neighbourhood in San Francisco?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup."

Your response were on average 120.03 items.

flights: In a first step, I had asked 100 times: "What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who have last year booked five intercontinental business class trips?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup."

Your response were on average 117.70 items.

techie: In a first step, I had asked 100 times: 'What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who often buy the latest technology very shortly after it has been released?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup.'

Your response were on average 115.05 items**.**

business: In a first step, I had asked 100 times: "What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who run an advertising agency?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price,

it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup."

Your response were on average 112.05 items.

oxford: In a first step, I had asked 100 times: "What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who hold an Oxford Executive MBA?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup."

Your response were on average 106.45 items.

pricesite: In a first step, I had asked 100 times: "What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who often go to price checking websites?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup."

Your response were on average 85.35 items.

ryanair: In a first step, I had asked 100 times: "What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who have last year booked two long-distance flights with Ryanair?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup."

Your response were on average 85.35 items.

suburb: In a first step, I had asked 100 times: "What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who live in a suburb that has lost a third of its population over the last decade?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup."

Your response were on average 76.51 items.

retired: In a first step, I had asked 100 times: "What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who are retired?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup."

Your response were on average 75.70 items.

kids: In a first step, I had asked 100 times: "What do you think: how many screens would the firm be able to sell at this price to customers with a credible interest in these screens who have three kids, and are the only breadwinner in the family?

For comparison please assume that there are enough potential customers in this subgroup. Hence if you estimate a number below 100, you effectively say that, at this price, it is harder to sell screens in this subgroup. If you estimate a number above 100, you effectively say that, at this price, it is easier to sell screens to this subgroup."

Your response were on average 76.15 items.

all treatments:
sales with a choice between three versions of the contract

The firm responds to the production challenges by an alternative offer:
a) "gold": delivery is guaranteed, with all specifications, at a defined date
b) "silver": delivery is guaranteed, with all specifications, but the firm reserves the right to adjust the originally indicated delivery date
c) "bronze": the firm reserves the right to cancel the contract

The website is programmed such that the firm only defines its price once the customer has fully specified the order, including the contract provisions.

What do you think: with the three options available, how many screens would the firm be able to sell to this subgroup of customers? If you estimate a number below **<specific number>** you effectively say that, when giving customers a choice between three different versions of the contract, it is harder to sell screens in this subgroup. If you estimate a number above **<specific number>**, you effectively say that, with contract choice, it is easier to sell screens to this subgroup.

In the next step, please estimate how many of the total number of sales will come with each of the three contract options. Please make sure that your estimates for "gold", "silver" and "bronze" add up to your estimate for "total".

Please use the following JSON format for the response:

```
{
  "estimated_sales_total": <number>,
  "estimated_sales_gold": <number>,
  "estimated_sales_silver": <number>,
  "estimated_sales_bronze": <number>
}”
```

II. Supplementary Experiment: User Prompt

(system prompt is the same as in the main experiment)

The firm would wish to exploit its first mover advantage; recall that currently the firm does not have a competitor. The firm expects willingness to pay for its custom screens to be quite heterogeneous. As screens are customized to the specific needs of every individual customer, the firm is not concerned that screens sold to one customer at a lower price might be resold, on a secondary market, to a second customer with a higher willingness to pay. Consequently the firm expects to make a considerably higher profit if it can estimate the willingness to pay of individual customers. However the firm only sells this one product. In the jurisdiction in which the firm operates, it is prohibited to sell cookie data. This is why the firm has no easy access to demographic markers. The firm considers using the design of the sales contract as a work-around. Rather than using the same standard form contract for all customers, it wants to give customers a choice. However the firm expects that customers would be overwhelmed by simultaneously choosing between multiple contract clauses. To make the choice salient, and in particular as diagnostic as possible about willingness to pay, the firm explores the following options:

a) cancellation of contract for exceptional increase in cost ruled out
b) guarantee of delivery date
c) test of functionality upon delivery
d) warranty period extended to 2 years
e) no use of purchase data in future negotiations

Can you please rank these options: how effective do you expect them to be as a diagnostic tool for willingness to pay?

Please explain your reasoning in natural language.

Please then summarize your ranking in JSON format, using only the characters a-e for the response. Hence if you believe that the most diagnostic clause would be an extension of the warranty to 2 years, your response should start with character "d", and so on.

Your final response should look as in the following example:

```
{
"ranking": "daebc"
}
```

III. Supplementary Results

1. Effect of Contract Choice Conditional on Chosen Contract

	profit	price	quantity
post	21785 (13436)	750.50*** (201.38)	-.520 (1.164)
savings	-88072*** (13436)	-934.50*** (201.38)	-11.62*** (1.164)
upgrade	224582*** (13436)	1925.50*** (201.38)	20.570*** (1.164)
silver	302616*** (13436)	3489.25*** (201.38)	21.540*** (1.164)
gold	197628*** (13436)	6520.50*** (201.38)	6.610*** (1.164)
post*silver	-6818 (19001)	-31.75 (284.79)	-1.610 (1.646)
post*gold	24623 (19001)	-213 (284.79)	3.020 (1.646)
savings*silver	-122254*** (19001)	-643* (284.79)	-4.820** (1.646)
savings*gold	281645*** (19001)	-445.50 (284.79)	38.330*** (1.646)
upgrade*silver	-311543*** (19001)	-644* (284.79)	-34.580*** (1.646)
upgrade*gold	-331130*** (19001)	-879** (284.79)	-33.160*** (1.646)
cons	815297*** (19568)	22159.50*** (142.4)	29.750*** (.823)
N	1200	1200	1200

Table 4
Effect of Contract Choice Conditional on Chosen Contract
OLS
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05

2. Effect of Demographic Information on Profit

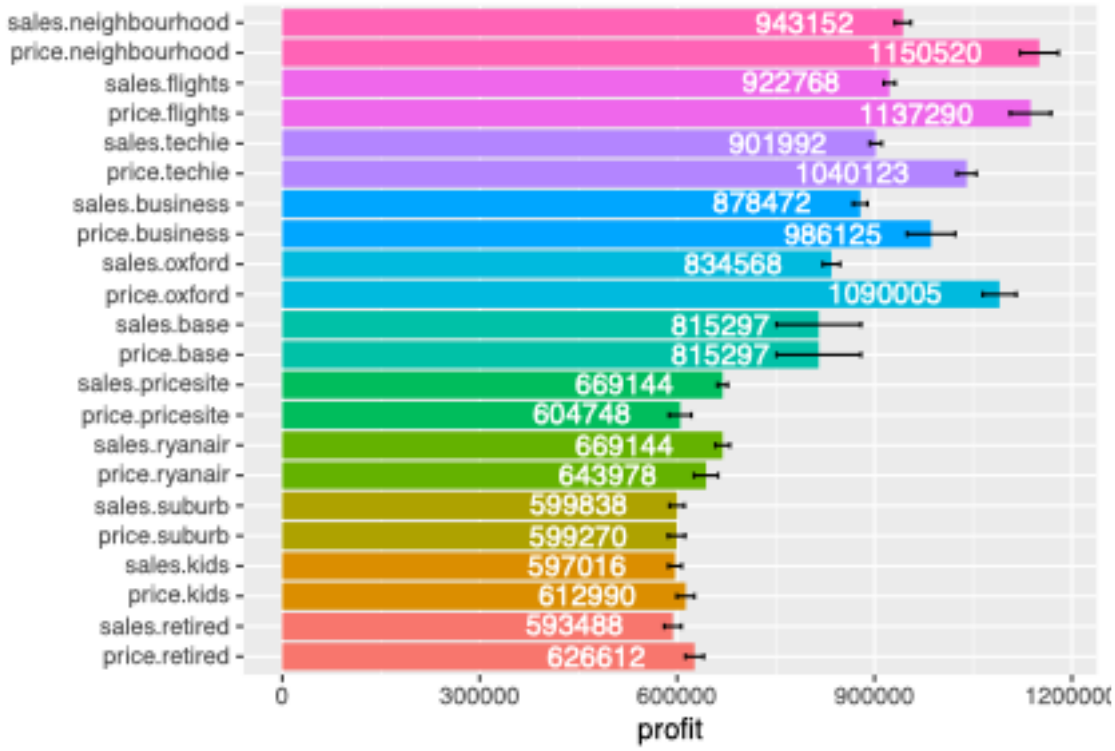

Figure 8
Alternative Estimations of Profit

	profp	profs
neighbourhood	335223*** (21004)	127855*** (15863)
flights	321993*** (21004)	107471*** (15863)
techie	224826*** (21004)	86695*** (15863)
business	170828*** (21004)	63175*** (15863)
oxford	274708*** (21004)	19271 (15863)
pricesite	-210549*** (21004)	-146153*** (15863)
ryanair	-171320*** (21004)	-146153*** (15863)
suburb	-216027*** (21004)	-215459*** (15863)
kids	-202307*** (21004)	-218281*** (15863)
retired	-188685*** (21004)	-221809*** (15863)
cons	815297*** (14852)	815297*** (11217)
N	1100	1100

Table 5
Profit in the Absence of Contract Choice
OLS
reference category: baseline
profp: calculations use estimated price, assuming 100 items are sold
profs: calculations use estimated quantity, assuming the price is 25000
standard errors in parenthesis
*** p < .001

3. Choice vs. Demographic Information

	profp
neighbourhood	19413 (17016)
flights	6183 (17016)
techie	-90984*** (17016)
business	-144982*** (17016)
oxford	-41102* (17016)
pricesite	-526358*** (17016)
ryanair	-487129*** (17016)
suburb	-531837*** (17016)
kids	-518117*** (17016)
retired	-504495*** (17016)
cons	1131107*** (12032)
N	1100

Table 6
Choice vs. Demographic Information
OLS
reference category: post
profp: calculations use estimated price, assuming 100 items are sold
standard errors in parenthesis
*** p < .001