Obfuscation maximization-based decision-making
Theory, methodology and first empirical evidence

Chorus, Caspar; van Cranenburgh, Sander; Daniel, Aemiro Melkamu; Sandorf, Erlend Dancke; Sobhani, Anae; Szép, Teodóra

DOI
10.1016/j.mathsocsci.2020.10.002

Publication date
2021

Document Version
Final published version

Published in
Mathematical Social Sciences

Citation (APA)
Chorus, C., van Cranenburgh, S., Daniel, A. M., Sandorf, E. D., Sobhani, A., & Szép, T. (2021). Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence. *Mathematical Social Sciences*, *109*, 28-44. https://doi.org/10.1016/j.mathsocsci.2020.10.002

Mathematical Social Sciences 109 (2021) 28–44

Contents lists available at ScienceDirect

Mathematical Social Sciences

journal homepage: www.elsevier.com/locate/mss

Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence

Caspar Chorus [a,*,1], Sander van Cranenburgh [a], Aemiro Melkamu Daniel [a], Erlend Dancke Sandorf [b], Anae Sobhani [c], Teodóra Szép [a]

[a] Department of Engineering Systems and Services, Delft University of Technology, Jaffalaan 5, 2628BX, Delft, The Netherlands
[b] Economics Division Stirling Management School University of Stirling, Stirling, UK
[c] Department of Human Geography and Spatial Planning in Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 8 January 2020
Received in revised form 18 August 2020
Accepted 6 October 2020
Available online 21 October 2020

Keywords:
Obfuscation
Signaling
Choice behavior
Preferences
Hiding

ABSTRACT

Theories of decision-making are routinely based on the notion that decision-makers choose alternatives which align with their underlying preferences—and hence that their preferences can be inferred from their choices. In some situations, however, a decision-maker may wish to hide his or her preferences from an onlooker. This paper argues that such obfuscation-based choice behavior is likely to be relevant in various situations, such as political decision-making. This paper puts forward a simple and tractable discrete choice model of obfuscation-based choice behavior, by combining the well-known concepts of Bayesian inference and information entropy. After deriving the model and illustrating some key properties, the paper presents the results of an obfuscation game that was designed to explore whether decision-makers, when properly incentivized, would be able to obfuscate effectively, and which heuristics they employ to do so. Together, the analyses presented in this paper provide stepping stones towards a more profound understanding of obfuscation-based decision-making.

1. Introduction

Models of rational decision-making are routinely based on the notion that agents base their choices on their latent, underlying preferences—and/or their goals, motivations, desires, needs[2]; see prominent examples from the fields of social psychology (Ajzen and Fishbein, 1977; Ajzen, 1991), behavioral decision theory (Edwards, 1954; Einhorn and Hogarth, 1981), mathematical psychology (Tversky, 1972; Swait and Marley, 2013), microeconomics (Samuelson, 1948; Houthakker, 1950; Sen, 1971), microeconometrics (McFadden, 2001; Walker and Ben-Akiva, 2002; Arentze and Timmermans, 2009; Marley and Swait, 2017), the decision sciences (Bell et al., 1988; Keeney and Raiffa, 1993), and artificial intelligence (Georgeff et al., 1998; Zurek, 2017). In other words, conventional models of decision-making routinely postulate that a decision-maker's latent preferences echo through in his[3] choices. It may even be said, that the notion that choices are signals of underlying preferences – as formalized in the revealed preference axioms – lies at the heart of most empirical work in decision-making; it is this assumption, which allows analysts to estimate preferences based on choice observations (e.g. McFadden, 1974, 2001; Small and Rosen, 1981; Ben-Akiva et al., 1985; McConnell, 1995; Train, 2009).

The decision-making model presented in this paper adopts a fundamentally different perspective, by postulating that in some situations, a decision-maker may wish to hide the preferences underlying his choices, from an onlooker. In other words, it captures the notion that the decision-maker may in some situations wish to *suppress* the echo of his preferences. The reasons for such obfuscation-based decision-making may include a decision-maker's wish to protect his privacy, or to avoid legal punishment or social shame. The proposed model of obfuscation-based decision-making is designed to be simple and tractable – it builds on the well-known concepts of Bayesian inference and information entropy – while still being able to capture subtle but important behavioral intuitions. In this paper, we will show

[*] Corresponding author.
E-mail address: c.g.chorus@tudelft.nl (C. Chorus).

[1] Except for first author, authors are listed in alphabetic order.

[2] We are aware that several scholars have made useful distinctions between these and related concepts and have ordered them in (cognitive) hierarchies. These distinctions and hierarchies are subject to considerable academic debate. In this paper we do not take a standpoint in this debate.

[3] For ease of communication, we refer to the decision-maker as "he" and to the onlooker as "she" throughout this paper, although either the agent and/or the onlooker may be conceived to be human or artificial ("it").

https://doi.org/10.1016/j.mathsocsci.2020.10.002
0165-4896/

that although the notion of obfuscation clearly goes against a fundamental premise underlying most decision theories, it is still possible to do meaningful normative and empirical analyses with a properly specified obfuscation model.

The notion of obfuscation-based decision-making is conceptually related to principal–agent interaction and mechanism design (Hurwicz, 1973), strategic ambiguity in political decision-making (Page, 1976; Kono, 2006), truth serums (Prelec, 2004), incentive compatibility (Carson and Groves, 2007), preference-falsification (Frank, 1996; Kuran, 1997), deception by artificial agents (Castelfranchi, 2000), privacy protection (Brunton et al., 2017) and covert signaling (Smaldino et al., 2018).

Despite this abundance of related work, this – to the best of the authors' knowledge – is the first paper to provide a model of the decision-making behavior of an agent that wishes to hide from an onlooker the latent underlying preferences that govern his choices. It is important to note at this point, that obfuscation – i.e., hiding preferences from onlookers – is fundamentally different from the much more widely studied notion of deception (e.g. Eriksson and Simpson, 2007; Van't Veer et al., 2014; Biziou-van-Pol et al., 2015; Danaher, 2020). We conceive deception in terms of an agent trying to mislead the onlooker into making her believe that a particular set of preferences underlies his choices while in reality, another set of preferences governed his decision-making. In contrast, an obfuscating agent has no 'target' set of preferences towards which he wants to steer the onlooker's beliefs; he merely wants to present the onlooker with as little as possible information regarding his preferences. Put colloquially: a deceiving agent wants the onlooker to give the wrong answer to the question "why did he do that?", while an obfuscating agent wants the onlooker to say "I do not know".

The remainder of this paper is structured as follows: Section 2 presents a model of obfuscation-based decision-making and illustrates some of its workings using a concrete, numerical example. Section 3 presents the results of an obfuscation game, designed to take a first step towards empirical validation of the obfuscation model. Section 4 concludes, and presents directions for further research. Four appendices are provided, which give important background information: Appendix A elaborates a number of decision-making contexts in which obfuscation is likely to be a preferred strategy for the decision-maker (flirtation in a bar, moral dilemmas, nuclear proliferation); Appendix B explores, using Monte Carlo analyses, the econometric identification of parameters in the obfuscation model; Appendix C presents the instructions as these were provided to participants to the obfuscation game, and Appendix D presents the choice tasks that were used in the game.

2. A model of obfuscation-based decision-making

In this section, we provide a formalization of the behavior of an obfuscating decision-maker. It is important to note, that in this section we do not yet adopt the perspective of an analyst focused on analyzing choices made by a set of decision-makers; in contrast, we focus on the behavior of an individual decision-maker; hence, we do not discuss any econometric considerations. Those will be the topic of Section 3.

Consider a decision-maker whose task is to choose an alternative from a set $\boldsymbol{A}$ containing J alternatives: $\left\{a_1 \dots a_j \dots a_J\right\}$. Set $\boldsymbol{G}$ contains K attributes (or goals, or criteria) on which the alternatives are assessed: $\{g_1 \dots g_k \dots g_K\}$. The extent to which the decision-maker cares about each particular attribute g_k is denoted by weights β_k. Assume for ease of communication, but without loss of generic applicability, that $\beta_k \in \{0, 1, 2, \dots, M\}\, \forall k$. That is, if the decision-maker does not care about a particular attribute, the associated weight equals zero; increasing values reflect increasing importance of the attribute; a weight of M reflects that the attribute is of the highest importance to the decision-maker. Scores x_{kj} which are stacked in a K by J matrix $\boldsymbol{X}$ reflect how each particular alternative scores on each particular attribute; the non-negative attribute-weights imply that higher scores are preferred over lower ones. The aggregated utility associated with choosing alternative a_j equals $u_j = \sum_{k=1}^{K} u_{jk}$, where $u_{jk} = \beta_k \cdot x_{kj}$. Note that this aggregation reflects a classical linear-additive multi-attribute utility approach; other aggregation procedures may be considered as well. Denote the K-dimensional vector containing the weights of all attributes as $\boldsymbol{\beta}$, which defines the decision-maker's preferences. The decision-maker's beliefs are defined as follows:

1. He is being watched by an onlooker.
2. The onlooker observes $\boldsymbol{A}$, $\boldsymbol{G}$, and $\boldsymbol{X}$; she has the same perception of these vectors and matrix as the agent himself.
3. The onlooker has uninformative prior probabilistic beliefs $P(\boldsymbol{\beta})$ about the weights attached by the agent to different attributes. She knows that each weight is an element from the set $\{0, 1, 2, \dots, M\}$. The onlooker's multidimensional uninformative prior thus consists of probabilities of size $1/(M+1)^K$ for each of the $(M+1)^K$ possible states of the world, where each state is characterized by a realization of each of the K weights β_k.
4. The onlooker observes one choice by the decision-maker from $\boldsymbol{A}$, and uses that observation to update her beliefs about weights $\boldsymbol{\beta}$, into posterior probabilities; she does so using Bayes' rule. Her posterior probabilities, after having observed the decision-maker's choice for alternative a_j, are given by:

$$P\left(\boldsymbol{\beta}|a_j\right) = \frac{P\left(a_j|\boldsymbol{\beta}\right) \cdot P\left(\boldsymbol{\beta}\right)}{\sum_{\boldsymbol{\beta}\in\boldsymbol{B}} P\left(a_j|\boldsymbol{\beta}\right) \cdot P\left(\boldsymbol{\beta}\right)} \tag{1}$$

Here $\boldsymbol{B}$ represents the domain of $\boldsymbol{\beta}$ (i.e., it contains all $(M+1)^K$ states of the world), and $P\left(a_j|\boldsymbol{\beta}\right)$ is given by the well-known Logit-formulation (Luce, 1959; McFadden, 1974) which stipulates that the probability of choosing an action given a set of preferences increases when the utility of that action (which is a function of the decision-maker's preferences and the action's scores) increases.

$$P\left(a_j|\boldsymbol{\beta}\right) = \frac{\exp\left(\sum_{k=1}^{K} u_{jk}\right)}{\sum_{l=1}^{J} \exp\left(\sum_{k=1}^{K} u_{lk}\right)} \tag{2}$$

In the following sub-sections, we will present a model of a 'preference-oriented' decision-maker who ignores the onlooker and only cares about making choices that are in line with his preferences; an 'obfuscation' agent who is only concerned with hiding his preferences from the onlooker; and a 'hybrid' agent who attempts to choose in line with his preferences while at the same time trying to avoid the onlooker learning those preferences. An illustrative example in the context of political decision-making is presented thereafter.

A 'preference-aligned' decision-maker applies his preferences to each alternative, giving:

$$u_j = \sum_{k=1}^{K} u_{jk} = \sum_{k=1}^{K} \beta_k \cdot x_{kj} \tag{3}$$

for alternative j; he then chooses the alternative with highest aggregated utility. An obfuscating decision-maker considers that the remaining uncertainty in the eyes of the onlooker, i.e. after having observed his choice for a_j, is quantified in terms of Shannon entropy (Shannon, 1948), where we use the decadic logarithm, without loss of generic applicability:

$$H_j = -\sum_{\boldsymbol{\beta}\in\boldsymbol{B}} \left[P\left(\boldsymbol{\beta}|a_j\right) \cdot \log\left(P\left(\boldsymbol{\beta}|a_j\right)\right)\right] \tag{4}$$

Table 1
States of the world and the onlooker's prior probabilities.

	$\beta_N = 0$	$\beta_N = 1$	$\beta_N = 2$
$\beta_E = 0$	(0,0) (1/9)	(0,1) (1/9)	(0,2) (1/9)
$\beta_E = 1$	(1,0) (1/9)	(1,1) (1/9)	(1,2) (1/9)
$\beta_E = 2$	(2,0) (1/9)	(2,1) (1/9)	(2,2) (1/9)

Table 2
Score matrix (political decision-making example: 2 attributes, 3 alternatives).

	a_1	a_2	a_3
s_E	3	1.5	0
s_N	0	1.5	3

The obfuscating agent chooses the alternative which maximizes entropy[4]: $\text{argmax}_{j=1..J}\ \{H_j\}$. A hybrid decision-maker's behavior is driven by a combination of preference-oriented behavior and entropy maximization, which may be represented by a utility-maximization process where the utility of an alternative is given as:

$$U_j = (1-\gamma)\cdot\frac{u_j - u_{\min}}{u_{\max}-u_{\min}} + \gamma\cdot\frac{H_j - H_{\min}}{H_{\max}-H_{\min}} \quad (5a)$$

or alternatively, without normalization, as:

$$U_j = u_j + \gamma \cdot H_j \quad (5b)$$

or alternatively, focusing purely on whether or not the considered alternative is the maximum entropy alternative, as:

$$U_j = u_j + \gamma \cdot \mathbf{1}\left(H_j \geq H_i\ \forall i \in C\right) \quad (5c)$$

In Eq. (5a), the utility of the most (least) attractive – in terms of preference-alignment – alternative in the set is denoted as $u_{\max}$ ($u_{\min}$). In Eq. (5c), indicator function $\mathbf{1}\left(H_j \geq H_i\ \forall i \in C\right)$ returns 1 if j generates more entropy than any other alternative in the choice set (C), and zero otherwise. Note that presumably these hybrid models (Eqs. (5a)–(5c)) have the strongest base in behavioral intuition: they represent a decision-maker who wishes to fulfill his preferences, but who is willing to give up some preference-related utility if this preserves his privacy in terms of prohibiting the onlooker to learn his preferences.

Appendix A provides several examples of real-world situations which may trigger obfuscation-based decision-making; in this Section, we consider and flesh out the following situation: a politician faces a public vote in favor of one of a set of 3 policy packages $\{a_1, a_2, a_3\}$ aimed at developing tourism in a region of great natural beauty. Each package is defined in terms of its economic (E) benefits and its protection of nature (N): $x_{Ej}, x_{Nj} \in [0, 3]$. Weights of attributes are $\beta_E, \beta_N \in \{0, 1, 2\}$. The decision-maker's utility function is given by Eq. (5b), where the preference-aligned part of utility equals $u_j = \beta_E \cdot x_{Ej} + \beta_N \cdot x_{Nj}$. Onlookers consist of colleagues in his political party as well as journalists. Their priors for the politician's attribute-weights are 1/9 for every state of the world (there being $3^2 = 9$ states of the world, as implied by a two-dimensional preference with three possible states for each dimension); see Table 1:

Score-matrix $\boldsymbol{X}$ is as follows (Table 2):

This score matrix reflects that policy package $\boldsymbol{a_1}$ scores very high on economic developments, but does nothing to protect nature; package $\boldsymbol{a_3}$ scores very high on nature preservation but fails to bring economic benefits; package $\boldsymbol{a_2}$ is a so-called compromise alternative (e.g., Simonson, 1989; Kivetz et al., 2004; Chorus and Bierlaire, 2013) which scores reasonably well on both preference-dimensions without attaining a stellar performance on either of them.

Before deriving which alternative is chosen by a politician who is interested in obfuscating his underlying attribute weights (preferences), let us first discuss why a politician might be tempted to obfuscate in the first place. Suppose that the politician's party, and society as a whole, is deeply divided on the issue (and that the politician knows this). He has a personal preference for economic benefits over environmental protection, but his main focus as a politician is to reduce migration flows into his country; as such he wants to avoid being drawn into a fight with either faction of his party (or with large shares of his constituency) over the tourism vote, also because he wants to save his political capital – e.g. in the form of bonds with political allies – to spend it on the migration topic which is much more dear to him. In such a situation, both a strategy of full transparency, dictating a vote for the package with greatest economic benefits, and a strategy of deception, dictating a vote for the package with maximum environmental protection, would cause problems for the politician in the sense that either option would suck him into a political fight that he wishes to avoid. An obfuscation strategy would make it difficult for onlookers to pinpoint – and subsequently attack – the politician's underlying political preferences; it would allow the politician to 'duck and cover', and move on to other political battles in which he is more interested.[5]

It is illustrative to derive first what the politician believes that the onlookers may learn – in terms of updating their flat priors into more informative posteriors regarding the politician's attribute weights – from his choice for a particular policy package. Applying Eqs. (1) and (2), Fig. 1 presents the onlooker's posteriors (to avoid repetition, we focus on alternatives $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$; note that alternative $\boldsymbol{a_3}$ is the mirror image of alternative $\boldsymbol{a_1}$).

In line with intuition, Fig. 1's top panel clearly shows that the politician's choice for package $\boldsymbol{a_1}$ (which scores high on Economy and low on Nature) results in the onlookers believing that the politician's weight for Economy is higher than that for Nature—i.e., states of the world (1,0), (2,0) and (2,1) become more likely, at the expense of states (0,1), (0,2) and (1,2) becoming less likely. The lower panel illustrates that a choice for compromise package ($\boldsymbol{a_2}$) which scores reasonably well on both preference-dimensions informs the onlookers that states (0,0), (1,1) and (2,2) are likely (each of these representing equal weights for both preference-dimensions) whereas states (0,2) and (2,0) (which imply that one attribute is much more important than the other one) become very unlikely. Also, this is in line with intuition. Note that a choice for alternative $\boldsymbol{a_3}$ generates posteriors that are the opposite of those generated by a choice for its mirror image alternative $\boldsymbol{a_1}$; that is, the posterior probabilities for states (0,0), (1,1) and (2,2) are the same as for $\boldsymbol{a_1}$, while those for $\boldsymbol{a_3}$'s state (1,2) equal the posterior for $\boldsymbol{a_1}$'s state (2,1), etc.

Eq. (3) uses these posteriors to give the Entropy associated with choosing a particular policy package: $H_1 = H_3 = 0.77$; $H_2 = 0.89$. It turns out that in this situation, choosing policy package $\boldsymbol{a_2}$, which represents the compromise option, is the optimal strategy for a politician who above all else wishes to avoid revealing his true preference-weights to onlookers. This finding

[4] But note that while the obfuscating agent chooses based on Entropy maximization, he is assumed to believe – see Eq. (2) – that the onlooker does not consider the possibility that he might obfuscate; in other words, he believes that the onlooker believes that his (the decision-maker's) choices are purely preference-aligned. At the end of this section, a more generic approach is formulated, which creates room for the possibility that the decision-maker believes that the onlooker does consider his obfuscation behavior.

[5] In the context of political decision-making, obfuscation is related to the concept of strategic ambiguity, although the latter notion does not involve inference of latent preferences from observed choices (e.g. Aragones and Neeman, 2000; Jarzabkowski et al., 2010). See also Kono (2006) for a study into the benefits of obfuscation in political decision-making. Jolink and Niesten (2020) provide evidence for the role of signaling in contexts where environmental and economic interest are being traded off against each other.

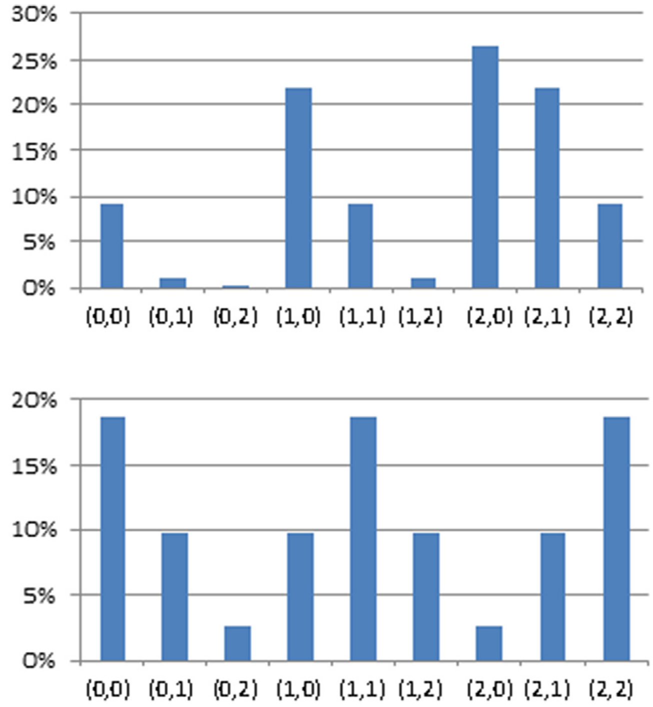


Fig. 1. Onlookers' posterior probabilities, having witnessed the politician vote for a particular policy package (top panel: $\boldsymbol{a}_1$, lower panel $\boldsymbol{a}_2$).

is in line with intuition: since a compromise option by definition scores reasonably well on each dimension, a choice for that option carries limited information about the weights attached by the decision-maker to different dimensions (compared to a choice for an option with more extreme performances on different dimensions). Note that this result contributes to the literature on compromise effects in Marketing, Transportation, Sociology, Decision-making and other fields (see earlier cited papers): compromise alternatives are known to attract disproportional demand, and various reasons have been put forward to explain this phenomenon; a wish to obfuscate is a new potential (partial) explanation for such an effect, particularly in political contexts such as the one described in this example.

Crucially, it depends on the politician's true attribute-weights whether or not obfuscation is costly to him: in case $\beta_E = \beta_N$, the politician derives an equal amount of 'preference-aligned utility' (i.e., $u_j = \beta_E \cdot s_{Ej} + \beta_N \cdot s_{Nj}$) from either policy package in Table 2. Therefore, in that case choosing the obfuscation option is costless to the politician—i.e., it leads to no loss in u. However, if the politician's true set of attribute-weights were $\beta_E = 2, \beta_N = 1$, choosing the obfuscation alternative ($\boldsymbol{a_2}$) instead of package $\boldsymbol{a_1}$ would lead to a loss in preference-aligned utility of size 1.5. Whether or not the politician is willing to give up this amount of preference-aligned utility to increase the onlookers' Entropy from 0.77 to 0.89 depends on the relative importance – i.e., γ in Eq. (5b) – which he attaches to obfuscation.

As a side note: which policy package is chosen by a *deceiving* (rather than obfuscating) politician depends on his true attribute-weights and what he wants the onlookers to believe. For example, if the politician's true weights are $\beta_E = \beta_N = 1$, and when he wants to make the onlookers believe that he cares less about the environment than about economic benefits, a choice for $\boldsymbol{a_1}$ would be the optimal strategy. This strategy boils down to costless deceit, as the preference-aligned part of utility is the same for each alternative in the set. However, if the politician's true weights were $\beta_E = 2, \beta_N = 1$, a choice for alternative $\boldsymbol{a_3}$, aiming to deceive the onlookers, would constitute costly deceit as the preference-aligned part of utility would be 3 units lower than for alternative $\boldsymbol{a_1}$. Again, it becomes clear that obfuscation and deception are two very different phenomena, implying different choices made by decision-makers.

At this point, it is important to return to the assumption, embedded in Eq. (2), that the decision-maker is assumed to believe that the onlooker presumes that he does not obfuscate. Inspired by the theory of cognitive hierarchy games pioneered by Camerer et al. (2004), we call the decision-maker under this assumption a Level 1 thinker who presumes that the onlooker is a Level 0 thinker. With this we mean that the decision-maker presumes that he is one step ahead, mentally speaking, of the onlooker who falsely believes that the decision-maker does not obfuscate, while he does. This assumption can be rationalized in several ways, e.g. by pointing at potential over-confidence (cognitive arrogance) from the side of the decision-maker: he may believe that the onlooker is cognitively unable to process and optimally respond to his obfuscation behaviors. But note that even if the decision-maker believes that the onlooker is equally smart as he is, he (the decision-maker) might still rightfully anticipate that it would be more difficult for her (the onlooker) to incorporate his obfuscation into her beliefs, than it is for him to obfuscate. Another possible justification for this Level 1–Level 0 discrepancy could be that the decision-maker believes that the onlooker is unaware of his wish to obfuscate. Such a belief may well be justified in situations where obfuscation is not an obvious strategy, or where the onlooker is considered naïve by the decision-maker.

In line with the thinking behind cognitive hierarchy games, we can relax this assumption of Level 1–Level 0 behavior in cases where it seems less justified, while maintaining, as per cognitive hierarchy theory, the assumption that the decision-maker is one step ahead of the onlooker. Let us consider the situation where the decision-maker obfuscates, by using Eqs. (5a), (5b), or (5c) with strictly positive obfuscation parameter γ. Now, rather than assuming that he (the decision-maker) believes that the onlooker is unaware of, or unable to process, his obfuscation behavior, we may assume that the decision-maker believes that the onlooker does consider the decision-maker's obfuscations as part of her thought process. This creates a Level 2–Level 1 constellation, which can be modeled by rewriting the probabilities of actions conditional on preferences, as presented in Eq. (2), into the following three variations, depending on the presumed type of (hybrid) obfuscation behavior:

$$P'\left(a_j|\boldsymbol{\beta}\right) = \frac{\exp\left((1-\gamma)\cdot\frac{u_j-u_{\min}}{u_{\max}-u_{\min}} + \gamma\cdot\frac{H_j-H_{\min}}{H_{\max}-H_{\min}}\right)}{\sum_{l=1}^{J}\exp\left((1-\gamma)\cdot\frac{u_l-u_{\min}}{u_{\max}-u_{\min}} + \gamma\cdot\frac{H_l-H_{\min}}{H_{\max}-H_{\min}}\right)} \tag{6a}$$

$$P'\left(a_j|\boldsymbol{\beta}\right) = \frac{\exp\left(u_j + \gamma\cdot H_j\right)}{\sum_{l=1}^{J}\exp\left(u_l + \gamma\cdot H_l\right)} \tag{6b}$$

$$P'\left(a_j|\boldsymbol{\beta}\right) = \frac{\exp\left(u_j + \gamma\cdot\mathbf{1}\left(H_j \geq H_i\ \forall i\in C\right)\right)}{\sum_{l=1}^{J}\exp\left(u_l + \gamma\cdot\mathbf{1}\left(H_l \geq H_i\ \forall i\in C\right)\right)} \tag{6c}$$

That is, the decision-maker that uses any of these probabilities acknowledges that the onlooker takes his inclination to obfuscate into account when she updates her probabilistic beliefs about his weights for the attributes (the βs embedded in u_j and u_l). The prime in $P'\left(a_j|\boldsymbol{\beta}\right)$ stands for the onlooker's place ('Level') in the cognitive hierarchy, as perceived by the decision-maker. The absence of a prime stands for Level 0, as in the original model, while the presence of one prime stands for Level 1, and so forth. Clearly, this changes posteriors from $P\left(\boldsymbol{\beta}|a_j\right)$ to $P'\left(\boldsymbol{\beta}|a_j\right)$ which implies that the resulting entropy changes as well. We denote this new entropy as H'_j. The decision-maker chooses based on H'_j, that is he chooses by means of maximizing utility over alternatives j, defined as either:

$$U_j = (1-\gamma)\cdot\frac{u_j - u_{\min}}{u_{\max} - u_{\min}} + \gamma\cdot\frac{H'_j - H'_{\min}}{H'_{\max} - H'_{\min}} \tag{7a}$$

$$U_j = u_j + \gamma\cdot H'_j \tag{7b}$$

$$U_j = u_j + \gamma\cdot\mathbf{1}\left(H'_j \geq H'_i\ \forall i \in C\right) \tag{7c}$$

We assume here, that consistent pairs are made between (6a) and (7a), and so forth. The fact that the decision-maker believes that the onlooker only takes into account the original entropy H_j while he uses H'_j in his decision-making, embodies the prevailing cognitive hierarchy in the eyes of the decision-maker, where both decision-maker and onlooker have now moved one step upwards in the hierarchy. In other words, the decision-maker still considers himself to be either smarter or less naïve than the onlooker (or he may believe that they are equally smart and cunning, but that processing obfuscating is more difficult from the onlooker's position). As described in Camerer et al. (2004), such moving up in the hierarchy can be iteratively continued until both the decision-maker and (his perception of) the onlooker are so sophisticated that the decision-maker's degree of obfuscation has become *de facto* common knowledge, at which point the game in the heads of the decision-maker and onlooker has reached an equilibrium. Empirical data of appropriately specified game-experiments can be used to estimate the relative position of the decision-maker/obfuscator and onlooker in the cognitive hierarchy, as we will show in the next section.

3. Empirical analysis based on an experimental economics approach

3.1. The obfuscation game

As a first step towards empirically validating the obfuscation model, an obfuscation game was developed in the tradition of experimental economics and induced value theory (Smith, 1976; Davis and Holt, 1993; Loewenstein, 1999; Kagel and Roth, 2016). That is, participants to our experiment were confronted with a carefully designed monetary incentive structure.

Incentives were designed such, that obfuscation was the optimal strategy for decision-makers playing the game. This way, by analyzing choice behavior of players, we were able to explore if – when properly incentivized – decision-makers would be able to identify and select the obfuscation option from a set of choice alternatives. This approach does not aim to explore if people obfuscate in real life or in experimental conditions that aim to mimic real life circumstances, but it rather tests the innate ability of people to obfuscate effectively, in case obfuscation behavior is the optimal decision-making strategy. As such, this induced value approach serves only as a very first step towards establishing empirical validation of the obfuscation model.

The goal of the obfuscation game is twofold: firstly, it aims to test whether or not, and to what extent, decision-makers succeed in identifying and selecting from a choice set the alternative which gives an onlooker minimum information regarding the motivation that underlies their choice. Secondly, it aims to explore what types of heuristics and/or cognitive processes are used by decision-makers in their attempts to obfuscate; this includes an empirical investigation into which cognitive hierarchy (see the discussion at the end of the previous section) has presumably driven the decision-makers' and onlookers' choice behavior.

To keep the experiment as tractable and understandable as possible, we chose to make two simplifications of the model presented in Section 2. First, we base the game on a situation where a decision-maker only considers one out of a number of attributes on which the alternatives are scored, as opposed to considering many attributes simultaneously, with different weights for each attribute. Second, each alternative could have one out of the following three 'scores' on each attribute: either the alternative is forbidden, allowed, or obliged by the attribute. In other words, the attributes can be thought of as rules; it is up to the onlooker to identify, based on the observed choice made by a decision-maker, which rule is followed by him. A particular challenge that has to be confronted in the design of the game's incentive structure, is that we want to clearly distinguish obfuscation ('hiding') from deception ('misleading'): the two notions are obviously related, as was discussed above, but the incentive structure needs to be designed in such a way that obfuscation behavior is optimal for the decision-maker, while deception is not. The solution for this challenge is found by designing an incentive structure in which:

1. the decision-maker receives money when the onlooker does not dare to guess the decision-maker's rule after he has made a choice from the set of alternatives;
2. the decision-maker receives no money when the onlooker attempts to guess his rule, irrespective of whether she guesses correctly or not;
3. the onlooker receives money when she refrains from guessing the decision-maker's rule;
4. the onlooker receives more money (compared to the previous bullet) when she correctly guesses the decision-maker's rule;
5. the onlooker receives no money when she incorrectly guesses the decision-maker's underlying rule.

The second feature of this pay-off structure allows us to distinguish between obfuscation and deception, and to rule out the latter: the decision-maker gains nothing from misleading the onlooker (i.e., trying to make her guess wrongly), and only gains from keeping her sufficiently 'in the dark' as to his underlying rule (such as to prevent her from guessing); this is exactly what obfuscation is about, and how it distinguishes itself from deception.

After an elaborate series of small-scale pilot studies in which we observed people play various specifications of the obfuscation game under slight variations of the incentive structure, the exact specifications were chosen as follows (note that the full instruction as read out to participants can be found in Appendix C; the choice cards are given in Appendix D):

- The game is played in decision-maker-onlooker pairs. Alternatingly, a participant plays the role of decision-maker and of onlooker. Each pair jointly executes ten choice tasks.
- Decision-maker and onlooker are informed about the payoff structure for both roles:
- Choice tasks take the form of a 5 by 5 matrix which is visible to both the decision-maker and the onlooker. Its rows represent rules, columns represent alternatives. An alternative is either obliged (!), allowed (✓) or prohibited (X) by a rule. Note that the designs of the choice tasks for the obfuscation game were made using the R-package *obfuscatoR* (Sandorf et al., 2019). The package was developed by the authors specifically to test the obfuscation hypothesis. Fig. 2 presents an example:
- For each choice task, the decision-maker is informed which rule he must follow (these differ per task; for the task depicted in Fig. 2, R1 must be followed). This information is not visible to the onlooker. Upon reading this private information and inspecting the publicly visible choice card, the decision-maker chooses an alternative which is compatible with his rule. In every choice task, there would always be three alternatives (out of five) which are compatible with

Choice task A					
Rules	Alternatives				
	A1	A2	A3	A4	A5
R1	✓	X	✓	X	✓
R2	X	!	X	X	X
R3	✓	X	✓	✓	X
R4	✓	✓	X	X	✓
R5	✓	✓	✓	X	✓

Choice task A; Rule to follow: R1

Fig. 2. Example choice task.

Onlooker / **Decision-maker**	**Does not guess**	**Guesses (correct)**	**Guesses (wrong)**
Chooses action allowed by rule	(10,5)	(0,10)	(0,0)

Fig. 3. The obfuscation game pay-off structure in normal form.

the decision-maker's rule, together forming his actual choice set. In the example of Fig. 2, these alternatives are A1, A3, and A5.

- After having indicated to the onlooker his chosen alternative, the onlooker chooses whether or not to guess the decision-maker's rule. If she chooses to guess, she tells the decision-maker which rule she believes governed his behavior.
- Irrespective of whether or not the onlooker guesses, the decision-maker subsequently informs her of his actual rule by showing the card on which it was written.
- The decision-maker and the onlooker jointly register this outcome on the payment form, which is updated by them after each choice task. The decision-maker receives 10 euro if the onlooker did not guess; 0 euro otherwise. The onlooker receives 5 euro if she refrained from guessing; 10 euro if she guesses correctly; 0 euro otherwise.
- After all ten rounds have been played, a plenary lottery is organized to draw two (out of ten) choice tasks: one in which the participant was a decision-maker, one in which the participant was an onlooker. The monetary outcomes associated with these drawn choice tasks are added to the fee of participating (which was 15 euro), and paid to the individual.

To summarize, Fig. 3 presents the pay-off structure of the game in normal form.

3.2. Data collection

On Thursday December 6th 2018, the game was played by 62 students (i.e., 31 pairs); they were recruited from among the 120 students taking an MSc-course "Statistical analysis of choice behavior". These students had recently obtained knowledge of choice modeling and discrete choice theory, but did not have any prior knowledge about the notion of obfuscation and how it could be modeled or computed, neither did they have knowledge about the concept of information entropy.[6] Participation was on a voluntary basis; it was made clear to participants, that neither participation nor performance in the game would in any way influence their grade for the course. Moreover, students were informed that their personal information would be unavailable to the lecturer of the course (being the first author of this paper); hence, the lecturer would not know who played the game and how well. Informed consent forms were provided to (and signed by) students before the start of the game, and the game itself was approved by the university's Research Ethics Committee; all relevant documentation can be obtained by emailing the first author. Playing the game took exactly one hour, including reading out the instructions, which were also available on paper for each pair. All participants who started the game, completed it; the average pay-off was 28.10 euro (which includes a 15 euro participation fee), with a minimum of 15 euro and a maximum of 35 euro; note that these were also the theoretical minimum and maximum.

3.3. Empirical analysis

Before presenting and interpreting the results obtained through the obfuscation game, it should once again be noted up front, that – in light of the game's controlled nature and the limited size of the sample – these analyses should only be considered a very first step towards empirical validation of the obfuscation model.

We start by analyzing to what extent participants succeeded in obfuscating, i.e., in selecting the alternative whose information entropy was highest within the choice set of feasible alternatives. Note that, given the two simplifications mentioned in 4.1, the process of Entropy maximization can be formalized as follows: consider an agent whose task is to choose an alternative from a set $\boldsymbol{A}$ containing 5 alternatives. Set $\boldsymbol{R}$ contains 5 rules, one of which the agent is assigned to follow. Matrix $\boldsymbol{S}$ which is 5 by 5-dimensional contains scores x_{kj} describing how alternative a_j performs on rule r_k. These scores may take on the following

[6] Clearly, this group does not form a representative sample of the population at large; this is one more reason why our empirical results should only be considered only a first step towards validation of the theory of obfuscation-based decision-making, which should receive further empirical scrutiny in larger, more representative follow up studies. For this reason, we chose not to register the usual socio-demographic attributes of participants.

values: $s_{kj} \in \{+, 0, -\}$. In case r_k is a so-called strong rule, $s_{kj} \in \{+, -\}$ implying that an alternative (or action) is either obliged (+) or prohibited (−) by the rule. In case r_k is a so-called weak rule, $s_{kj} \in \{0, -\}$ implying that an alternative is either permitted (0) or prohibited (−) by the rule. A strong rule can thus alternatively be seen as a weak rule with only one alternative being permitted. The agent's beliefs are as follows: he is being watched by an onlooker. The onlooker observes $\boldsymbol{A}$, $\boldsymbol{R}$, and $\boldsymbol{S}$, and has the same perception of these sets and matrix as the agent himself. The onlooker has uninformative prior probabilistic beliefs about which rule from $\boldsymbol{R}$ governs the agent's decision-making behavior. Specifically: $P(r_k) = 1/5$ for each rule.[7] The onlooker observes one choice by the agent from $\boldsymbol{A}$, and uses that observation to update her probabilistic beliefs about which rule from $\boldsymbol{R}$ is adopted by the agent, into posterior probabilities; she does so using Bayes' rule. Specifically, the onlooker's posterior probabilities, after having observed the agent's choice for a_j, are given by: $P\left(r_k|a_j\right) = \frac{P\left(a_j|r_k\right) \cdot P(r_k)}{\sum_{k=1}^{K}\left[P\left(a_j|r_k\right) \cdot P(r_k)\right]}$, where $P\left(a_j|r_k\right)$ is defined as follows: if r_k is a strong rule, then $P\left(a_j|r_k\right) = 1$ if a_j is obliged under r_k, that is, if $x_{kj} =$ '+'. Otherwise, $P\left(a_j|r_k\right) = 0$. If r_k is a weak rule, then $P\left(a_j|r_k\right) = 0$ if a_j is prohibited under r_k (i.e., if $x_{kj} =$ '−'.) and $P\left(a_j|r_k\right) = 1/L_k$ otherwise, where L_k equals the size of the subset $\mathcal{L}_k$ of alternatives permitted under r_k. The obfuscating decision-maker considers that the remaining uncertainty in the eyes of the onlooker, i.e. after having observed his choice for a particular alternative a_j, is quantified as: $H_j = -\sum_{k=1}^{K}\left[P\left(r_k|a_j\right) \cdot \log\left(P\left(r_k|a_j\right)\right)\right]$.

Take the example of choice task A (as in Fig. 2), where alternatives A1, A3, and A5 are allowed by the decision-maker's rule (R1); following the above model, A1's entropy equals 0.6, whereas that of A3 and A5 equals 0.47. Thus, in the context of this choice task, and given the decision-maker's rule-assignment, alternative A1 is the obfuscation alternative as it generates maximum entropy for an onlooker with uninformative priors. Each choice task was designed in such a way, that there would always be one alternative whose entropy was higher than that of all other alternatives available to the decision-maker—that is, there would always be one alternative whose selection would be optimal for an obfuscating decision-maker.

Results obtained from analyzing the choices made by decision-makers are encouraging, when it comes to their capacity to identify the maximum entropy alternative from the choice set. To start with, in nine out of ten choice tasks, the maximum entropy alternative had the highest 'market share' of the three alternatives in the choice set of feasible alternatives. The exception is choice task 10 (see Appendix B), where alternative 4 had a lower entropy than the highest-entropy alternative 5 (0.41 versus 0.48) but was slightly more often chosen by decision-makers (15 times versus 13 times). Furthermore, in nine out of ten choice tasks, the lowest entropy alternative – that is, the alternative which most clearly gives away the decision-maker's rule – had the lowest market share. The exception being choice task 7 (see Appendix B), where alternative 3 has the lowest entropy (0.3) and was chosen by 4 decision-makers, whereas alternative 2 has a somewhat higher entropy (0.41) but was chosen by 3 decision-makers. In 59% of choices made by decision-makers (i.e., 193 out of 310), the maximum entropy alternative was chosen; this should be compared with a benchmark of 33% random chance given the size of the feasible choice set, which as mentioned earlier contained three alternatives in each choice task. Only in 9% of cases (29 out of 310), did the decision-maker select the minimum entropy alternative—this too, should be benchmarked against the chance probability of 33%. For a more detailed analysis, we created an entropy index, which assigns the value 0 to the entropy of the minimum entropy alternative in a particular choice set of feasible alternatives and the value 1 to the entropy of the maximum entropy alternative in that set. Using this index, we find that the mean index-value associated with the decision-maker's choice equals 0.80; colloquially, this implies that on average, decision-makers succeeded in generating 80% of the potential entropy that is 'available' to them in a choice task.

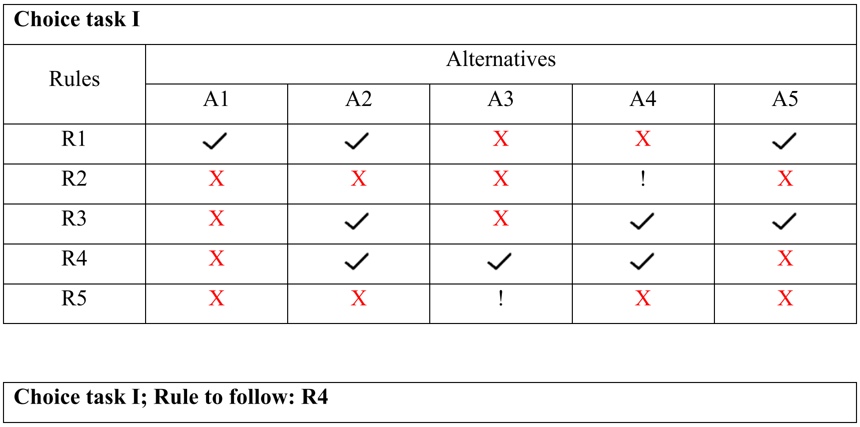
Given this fairly successful obfuscation behavior exhibited by decision-makers, it comes as no surprise that the onlooker in most cases did not dare to guess the decision-maker's underlying rule, although it is likely that risk aversion has also played a role, noting that for not guessing the onlooker could earn 5 euro easily. More specifically, only in 17% of cases (51 out of 310) did the onlooker guess the decision-maker's rule; and only in 37% of those cases (19 out of 51), did she do so correctly. This further corroborates our finding that participants to the experiments succeeded quite well in hiding their rules from the onlooker.

Following our assessment of the extent to which participants succeeded in obfuscating their rules, we now attempt to answer the question *how* they obfuscated, i.e., which heuristic, if any, was used. It should be noted, that we refrained from asking decision-makers directly how they arrived at their choices, thus relying solely on observed choice patterns to identify and compare heuristics. This is in line with the well-established notion, that people's explanations of why and how they arrived at certain decisions tend to be unreliable post-hoc rationalizations, offering little insight into actual decision processes (Nisbett and Wilson, 1977; Haidt, 2001). We distinguish between two heuristics, and we compare them with the sophisticated mechanism assumed in the obfuscation model (based on Bayesian learning and entropy maximization):

1. For each of the feasible alternatives, count the number of rules that support the alternative, and choose the alternative that is supported by the maximum number of rules. Note that for each choice task in the game, the maximum entropy alternative is also supported by the maximum number of rules, but in no less than eight out of ten choice tasks, following this heuristic fails to unambiguously identify the maximum entropy alternative (i.e., there would be a tie between two alternatives).
2. For each of the feasible alternatives, identify which rules support (i.e., oblige or permit) the alternative. For each of these rules, count the number of alternatives supported by this rule. Sum those numbers across rules, and maximize the outcome (over the feasible alternatives). Following this heuristic, which is more sophisticated than the previous one, always leads to unambiguous identification of the maximum entropy alternative in the context of the choice tasks used in our game.

The first heuristic is straightforward: it is based on the intuitive notion that when an alternative is supported by ('compatible with' or 'explainable in terms of') many rules, this makes it difficult for the onlooker to guess, having observed a choice for the alternative, which rule led to this choice. Take the choice task presented in Fig. 4 (which is choice task I as presented in

[7] As part of our empirical analyses, we tested this assumption of uniform priors versus an alternative assumption that a probability of (almost) zero is assigned to strong rules which only allow one action. The idea behind this latter assumption is that, although our instructions to participants made it clear that also strong rules were possible candidates to be assigned to decision-makers, such a rule would not leave the decision-maker with a choice between different actions and as such might have been considered an unlikely candidate rule by players of the game. Our empirical analyses – which can be obtained from the first author but are not reported here for reasons of space limitations – suggest that the assumption of uniform priors provides a closer fit with the observed choice data than its non-uniform counterpart.



Choice task I					
Rules	Alternatives				
	A1	A2	A3	A4	A5
R1	✓	✓	X	X	✓
R2	X	X	X	!	X
R3	X	✓	X	✓	✓
R4	X	✓	✓	✓	X
R5	X	X	!	X	X

Choice task I; Rule to follow: R4

Fig. 4. Example choice task.

Appendix D): here, alternatives A2, A3, and A4 form the feasible set, given the decision-maker's rule R4.

Alternative A2 is supported by 3 rules, A3 by 2 rules, and A4 by 3 rules. Their entropies are 0.48, 0.24, and 0.41 respectively. Following the heuristic "counting the number of supporting rules" would lead to a choice for either A2 or A4; indeed this subset contains the maximum entropy alternative (A2), but the decision-maker following this heuristic is left with a tie between the two alternatives. This tie can be avoided or broken, by following the second heuristic: alternative A2 is supported by three rules, which each support three alternatives, leading to a value of 9 for alternative A2. A similar counting exercise leads to the value 7 for A4 (and a value of 4 for A3). Maximization implies a choice for A2, which indeed is the maximum entropy alternative. The intuition behind this more sophisticated heuristic is as follows: obfuscation consists of making the link between an alternative and the rule which led to the alternative as unclear as possible. This can be done by maximizing the number of rules that support a particular alternative, but an additional factor that may be taken into account, is to ensure that those rules that support your alternative, also support as many as possible other alternatives. This additional aspect is captured in the second heuristic. Employing this second, more sophisticated heuristic highlights a key difference between A4 and A2: R2, being one of the rules which supports A4, supports no other alternative, while each of the rules supporting A2 supports several other alternatives as well. This makes that the onlooker has more difficulty guessing the underlying rule from A2 compared with A4.

Empirical analysis of our data show that in 90% of cases (278 out of 310), the chosen alternative had the highest number of supporting rules within the feasible choice set, possibly tied with another alternative. More specifically, in the two choice tasks where there was no such tie, 74% of choices was for the alternative that was supported by the maximum number of rules; this should be benchmarked against 33% random chance. In those eight choice tasks where two out of three feasible alternatives had the highest number of supporting rules (i.e., where following heuristic 1 leads to a tie), 94% of decision-makers selected one of the two alternatives with the highest number of supporting rules; this should be benchmarked against 67% random chance. These results suggest that heuristic 1 has helped participants in their search for the obfuscation – i.e., maximum entropy – alternative.

Regarding heuristic 2 – which, in addition to maximizing the number of rules supporting a particular alternative, also considers and maximizes how many alternatives are supported by each of those supporting rules – we find that in 59% of cases (183 out of 310), the selected alternative was compatible with this more sophisticated heuristic; this should be benchmarked against 33% random chance. Interestingly, in those cases where following heuristic 1 would lead to a tie between two alternatives, the alternative compatible with heuristic 2 was selected in 62% of cases (which should be benchmarked against 50% random chance), suggesting that heuristic 2 may in some cases have been used as a tie-breaker.

To refine our analysis beyond the descriptive statistics presented above, we estimated a series of Logit-models based on (combinations of) heuristics, see Table 3. For the decision-maker (DM), the unit of analysis is the multinomial choice for a particular alternative from the set of three feasible alternatives. Results can be summarized as follows, focusing first on models DM1-3: for the decision-maker, parameters are of the expected positive sign and they are all significant, signaling that an alternative's chance of being selected, increases if: the number of rules supporting the alternative increases (DM1); the summation, across rules supporting an alternative, of the number of alternatives supported by that rule, increases (DM2); the Entropy of the alternative increases (DM3). The Ben-Akiva and Swait (1986) test for non-nested models suggests that best-fitting model DM3 performs better than the second best-fitting model (DM1) with a p-value of 0.011.

Model DM4 is a variation of model DM3, the difference being that DM4 assumes a Level 2 – Level 1 hierarchical constellation while DM3 assumes a Level 1 – Level 0 hierarchical constellation (see the end of Section 2). In other words, DM4 is based on the assumption that the decision-maker believes that the onlooker is aware of the fact that he obfuscates; and that she (the onlooker) also processes this when deciding to guess or not. In light of the rules of the game played by the decision-maker and the onlooker, this assumption seems more realistic than the one embedded in DM3, which is that the decision-maker believes that the onlooker fails to process his obfuscation behavior in her decision whether or not to guess. In notation: DM3, as discussed earlier, assumes that the decision-maker believes that the onlooker believes that he would pick randomly from the actions allowed by his rule, denoted as $P\left(a_j|r_k\right) = 1/L_k$, where L_k equals the size of the subset $\mathcal{L}_k$ of alternatives permitted under r_k. In contrast, DM4 assumes that the decision-maker believes that the onlooker believes that he would choose from the actions allowed by his rule, according to a Logit model that assigns a high probability to actions whose entropy is high: $P'\left(a_j|r_k\right) = \frac{\exp(H_j)}{\sum_{l\in\mathcal{L}_k}\exp(H_l)}$. This leads to a different entropy H'_j compared to that of the initial model (H_j). While DM4 is still based on the notion that the decision-maker (believes he) is one step ahead of the onlooker, both have now climbed one level higher on the cognitive hierarchy ladder.

Estimation results suggest that the choices made by decision-makers in the obfuscation game are slightly better explained

Table 3
Estimation results.[7]

Agent	ID	Utility function	$\hat{\beta}$	SE($\hat{\beta}$) (rob.)	LL_0	$LL_{\hat{\beta}}$
Decision-maker	DM1	$V_j = \beta_R^{DM} \cdot R_j$	1.68	0.19	−340.6	–270.5
	DM2	$V_j = \beta_{RS}^{DM} \cdot RS_j$	0.41	0.04	−340.6	–285.7
	DM3	$V_j = \beta_H^{DM} \cdot H_j$	10.8	1.25	−340.6	–267.3
	DM4	$V_j = \beta_H^{DM} \cdot H'_j$	10.1	1.15	−340.6	–266.8
Onlooker	O1	$V_g = \beta_R^O \cdot R_{DM}$	−1.01	0.39	−214.9	–134.5
	O2	$V_g = \beta_{RS}^O \cdot RS_{DM}$	−0.25	0.09	−214.9	–134.7
	O3	$V_g = \beta_H^O \cdot H_{DM}$	−3.66	0.33	−214.9	–134.8
	O4	$V_g = \beta_H^O \cdot H'_{DM}$	−3.67	0.47	−214.9	–137.3

by model DM4 than DM3, suggesting that cognitive hierarchy constellation Level 2 – Level 1 fits the data slightly better than constellation Level 1 – Level 0, but the difference in final log-likelihood is too small to attach much certainty to this finding: the Ben-Akiva and Swait test suggest a p-value associated with the difference in model fit equaling 0.159. This implies that the difference in model fit is only significant at a modest 10%-level, if a one-tailed test is applied based on the notion that the rules of the game, which were common knowledge, make it reasonable to expect that decision-makers anticipate that onlookers take into account their (the decision-makers') obfuscation in their own decision-making processes.

Finally, we estimate models of onlooker behavior. Here, the unit of analysis is a binary choice to guess (denoted g in Table 3) or not, having been presented with the alternative selected by the decision-maker. Note that constants were estimated, but found to be far from significant, and left out of the final models. Results can be summarized as follows. Again, we focus first on models O1–O3: for the onlooker, parameters are of the expected negative sign and they are all significant, signaling that the onlooker's probability of guessing decreases if: the number of rules supporting the alternative chosen by the decision-maker increases (O1); the summation, across rules supporting the alternative chosen by the decision-maker, of the number of alternatives supported by that rule, increases (O2); the Entropy of the alternative chosen by the decision-maker increases (O3). The Ben-Akiva and Swait (1986) test for non-nested models suggests that best-fitting model O3 does not perform significantly better than the second best-fitting model (O1); the corresponding p-value equals 0.081. Comparing the log-likelihoods of models O4 and O3, it appears that there is no evidence for the assumption, embedded in model O4, that the onlooker takes into account that the decision-maker anticipates that she (the onlooker) takes into account his (the decision-maker) obfuscation behavior.

To sum up: results suggest that in the obfuscation game, both players' behavior fits a Level 2–Level 1 cognitive hierarchy wherein the onlooker takes into account the decision-maker's obfuscation behavior (as opposed to presuming that he selects actions at random), while the decision-maker takes into account this awareness from the side of the onlooker, as such remaining one step ahead of her.

4. Conclusions and directions for further research

This paper puts forward a model that is based on the postulate that decision-makers in some situations may wish to hide the latent preferences governing their observable choices from an onlooker. As elaborated in Appendix A, such obfuscation-based behavior may be relevant in various agent–onlooker interactions. The paper presents a model that is rich enough to capture important yet subtle intuitions regarding obfuscation-based decision-making (and to clearly distinguish obfuscation from deceit), while maintaining a high level of tractability. After discussing and illustrating the workings of the model, and elaborating how it can be framed in the tradition of cognitive hierarchy games (Camerer et al., 2004), we present the results of an obfuscation game that is developed in the tradition of experimental economics. Results of this first step towards empirical validation of the obfuscation mode can be summarized as follows: when properly incentivized, participants are rather successful in identifying and selecting from a choice set the obfuscation alternative which generates maximum entropy to an onlooker. And: obfuscation-based decision-making behavior tends to align with simple heuristics, but there is also evidence of more sophisticated considerations by decision-makers. In particular, our findings suggest that a cognitive hierarchy was present where the onlooker anticipated obfuscation behavior from the side of the decision-maker, while the decision-maker by taking this into account stayed one step ahead of the onlooker.

In the process of designing a tractable obfuscation model, trade-offs were made, which we will not obfuscate but rather highlight, as they may provide useful starting points for further research: to start with, we focused on a one-shot application, where the decision-maker chooses an alternative from a set once. A natural model extension would be to consider a repeated choice situation. Related to this, we have focused on *decision-maker* behavior only, whereas future research may also consider (active) behavior by the onlooker. For example, the onlooker may be given the task to design a choice set for the decision-maker to choose from. In a repeated choice setting, onlookers and decision-makers will then interact in terms of providing choice sets (the onlooker) and choosing from those sets (the decision-maker). In such a model, the attribute weight-posteriors obtained by the onlooker in one choice situation may be used as attribute weight-priors for the next one.

A related direction for further theoretical research would be to relax the assumption that onlooker and agent share the same knowledge concerning the set of attributes, the set of alternatives, and the score of each alternative on each attribute. More generally, the models proposed in this paper can be extended by relaxing their underlying assumptions regarding, for example, the number of attributes and alternatives in the choice set (what happens to obfuscation behavior when attributes or alternatives are added to or removed from the set?) and the updating process including our use of an uninformative prior (what happens when other update processes are considered, and when priors are based on previous experience and hence not completely uninformative?). Studying such adaptations are worthwhile directions for further research.

[7] Note that we also tested various combinations of heuristics, as well as latent class models (each class representing a different heuristic) for both the decision-maker as the onlooker, but unsurprisingly, those models led to highly correlated estimates and no improvements in model fit, reflecting the intrinsic difficulty of distinguishing subtly different (obfuscation) heuristics based on observed choice patterns alone.

In a more general sense, one could argue that this paper puts much weight on the conceptual introduction of obfuscation maximization and on how to model this as a behavioral phenomenon, introducing only a limited degree of formalization. In future work, the obfuscation model should be embedded within a more axiomatic and rigorous formal framework, which for example would elaborate under which conditions obfuscation is a rational (optimal) decision strategy. Strong contenders for frameworks which would allow for such meta-reasoning are game theory, e.g. the use of repeated Von Stackelberg games (Von Stackelberg, 2010), and the belief–desire-intention formalism used in the artificial intelligence ('multi-agent systems') community (Georgeff et al., 1998). We consider the development of such improved formalizations to be core avenues for future research.

Furthermore, one could see obfuscation-based decision-making as a special case of a more general class of information regulation models, which presume that decision-makers are aware of, and actively manage, the amount of information concerning their preferences which is signaled through their choices to observers. The opposite extreme, and another special case of such information regulation behavior, is the notion of full transparency, where a decision-maker makes choices that provide purposely clear signals about his preferences, e.g. to signal his morality or social status. Such entropy-*minimization* behavior could be linked to the well-known phenomenon of 'conspicuous consumption' (e.g., Bagwell and Bernheim, 1996).

As a final note, although our empirical results can be considered promising, it is important to again highlight that they provide only very first steps towards validation and exploration of obfuscation-based decision-making. Aside from the usual caveats relating to experimental economics work, we must also mention here that the careful design of the incentive structure (including the use of small-scale pilots) could have inadvertently led to so-called forking, which increases the likelihood of finding statistically significant effects due to pure chance (Gelman and Loken, 2013). Crucially, follow up research would need to consider more real-life situations and larger, more representative samples, moving from the realm of experimental economics (where preferences and obfuscation mechanisms are induced by the analyst) to other tools for empirical data collection such as stated choice experiments and revealed choice data sets. Our Monte Carlo analyses provide some initial confidence that if obfuscation behavior is present in such data, a properly specified choice model could be able to retrieve it in a process of maximum likelihood estimation. The analyses presented in this paper may thus serve as guidance for these important next steps in understanding and modeling obfuscation-based decision-making.

Acknowledgments

This research is sponsored by the European Research Council, by means of ERC-Consolidator grant 724431-BEHAVE. The first author is grateful for many thoughtful comments and suggestions received during Q&As following keynotes at the STRC conference (Ascona) and the Netherlands Ministry of Foreign Affairs; presentation at the hEART conference (Athens) and the ICMC (Kobe, Japan); and seminars at TU Berlin, Leeds University, and TU Delft. Erlend Dancke Sandorf made his contributions while working at the Swedish University of Agricultural Sciences and gratefully acknowledges funding from the Jan Wallander and Tom Hedelius foundation, and the Tore Browaldh foundation, Sweden (Grant No. B2015-0490:1). We thank student–assistant Chris Van Bergen for her invaluable help in implementing and executing the obfuscation game–experiment. The data obtained from the experiment will be made freely accessible at https://data.4tu.nl/repository/. Comments made by two anonymous reviewers have greatly helped us improving a previous version of the paper, for which we are grateful.

Appendix A. Examples of situations that may trigger obfuscation-based decision-making

As will be made clear in a series of examples, there may be compelling reasons why, in certain situations, obfuscation may be more beneficial to the agent than either being transparent or deceiving the onlooker. Every example follows a similar line of reasoning: a situation is described in which an agent faces a choice from a set of alternatives. Some of these alternatives would give away his latent preferences, while other alternatives would either obfuscate or deceive the onlooker. It is then discussed why, in these particular situations, obfuscation may be the best strategy for the agent. Note that in these examples, we will use different terms (goals, preferences, principles, rules, etc.) for the latent construct that governs choices, depending on what makes most sense in the particular context.

A.1. Obfuscation in flirtation

To start on a relatively light-hearted note: consider the situation where the agent is having a drink in a bar, and a small group of friends enters the room. The agent has a romantic interest in one of the friends, and faces a choice from a set of alternative actions, including: whether or not to start a conversation with the group; with one of the group members in particular; offer one or all of them a drink; ignore them altogether, etc. A strategy of full transparency would dictate that the agent actively engages with the one group member whom he has a romantic interest in, immediately starting a conversation and perhaps offering him or her a drink. From such actions, the onlooker(s) would easily infer the preference of the agent. However, there may be several reasons why an agent would not want to use this strategy of full transparency, one compelling reason being that if the subject of his romantic interest turns out not to be interested in him, he would face public embarrassment.

A strategy of deception on the other hand would dictate that the agent could either choose actions that signal his lack of interest in his subject of interest or in any of the group's members (e.g. by ignoring them altogether), or choose actions that would signal his interest in *another* member of the group (e.g. by actively courting that other person). A clear disadvantage of such a deception strategy is that, while it could help avoid embarrassment, the chance that the agent will end up satisfying his romantic preference is small given this strategy. A strategy of obfuscation would dictate, that the agent acts in a way that on the one hand increases the likelihood of getting the positive attention of his subject of interest, while on the other hand reducing the probability of immediately and fully giving away his romantic interest and subsequently being embarrassed. One such obfuscation action would be to engage casually with the group as a whole, and gradually focusing attention towards the subject of interest, in case small positive signals are received from his or her side.

A.2. Obfuscation in a moral dilemma

Consider the situation where an agent is faced with a moral dilemma while being observed by his social peers. Specifically, each of the alternative actions available to the agent will violate important moral principles while adhering to other important moral principles. For example, in the situation where the agent has cheated on his partner, actions could include 'do everything you can to avoid your partner from finding out' and 'tell your partner what happened'. The former of these would prioritize the moral principle 'do not harm a loved one', while the latter would prioritize the moral principle 'do not lie to a loved one'. The agent anticipates that his actions are observable to his friends

– the onlookers – which, after having observed the agent's choice for a particular action, will use that choice to infer which moral principle has presumably guided his choice. The agent anticipates that based on this inference, some of the onlookers will 'punish' him with indignation, contempt or worse, if they believe that the wrong moral principle is prioritized. A strategy of full transparency (to his friends) would dictate that the agent fully aligns his action(s) with his guiding moral principle. This implies that, depending on his principle, either (and somewhat ironically) he makes a genuine attempt to make sure his partner does not find out, or that he tells his partner what happened. A strategy of deception would dictate that the agent deliberately tries to mislead his friends regarding his moral priorities: for example, in case the agent's true priority is not to lie to his partner, then he would mislead his friends if his actions would signal to them that his priority is not to harm his partner (e.g. by actively avoiding that his partner would find out).

For any of these strategies to work, the agent must first know his own, true moral priorities. However, it is well established that in many moral dilemmas, humans have a very hard time figuring out which moral principle should take priority (Forsyth and Nye, 1990; Sunstein, 2005; Gigerenzer, 2010; Capraro and Rand, 2018). In addition, the deception strategy can only work if the onlookers share one moral priority and if the agent knows this. In many situations, one or both of these two conditions will not be met, making deception an ineffectual or even impossible strategy. An obfuscating strategy may remedy this problem, by making it unclear to onlookers which moral principle has guided the agent's actions. One such action could be, not to actively inform one's partner but at the same time not to make active attempts to hide the cheating. By choosing this course of action, the agent can claim to adhere to both moral principles at once, or at least not to actively violate any of them.[8]

A.3. Obfuscation in nuclear non-proliferation

Consider the situation where a state wishes to keep its nuclear options open, in the sense that it wishes to create an ability to develop a nuclear weapon in the future—in case geopolitical developments would demand that. In this vein, the state pursues a program of technological developments that would enable it, if need be, to jumpstart the rapid development of a nuclear weapon. The international community, represented by the International Atomic Energy Agency (IAEA) of which the state is a member, audits such technological programs to ensure that no state other than those who already have nuclear weapons, develops them. Importantly, the IAEA does allow for the development of nuclear technology for non-military (e.g., energy) purposes, and some of the technology needed to develop nuclear weapons is so-called dual use: it can be used for either energy or military purposes. However, some of these dual use technologies are more effective for energy-related purposes, others being more effective for weapon development. The alternative actions available to the state agent are specific paths of technological development. The onlooker (IAEA) observes the actions chosen, and from them tries to infer whether the underlying motivation is energy- or military-related. A strategy of full transparency would imply that the state actor would choose the technological development-path which scores best on the goal 'prepare for future nuclear weapon development'. This would obviously trigger sanctions of the IAEA and potential geopolitical isolation. A strategy of deception would imply that the state actor would choose the technological development-path which scores best on the goal 'build a nuclear energy system'. This would clearly avoid sanctions, but at the same time it would not bring the state much closer to its true goal. An obfuscation strategy would go some way to help avoid both these disadvantages: it would imply acquiring those dual use technologies that score reasonably well on both goals, even if the technology is not the most effective one on either goal. Such a compromise makes it hard for the onlooker to learn the true goals of the agent, while the agent does not handicap himself in the process, by foregoing crucial technologies.

A.4. In sum: reasons to obfuscate rather than being transparent or deceiving

There are various reasons why an agent would be tempted to obfuscate (i.e. "create a smoke-screen") rather than simply giving away his latent goals by means of allowing an onlooker to easily learn them based on observing the agent's choice. The agent might be afraid that he will be punished if the onlooker learns his goals. This punishment may take the form of rejection (the flirtation example), contempt (the infidelity example), political damage (the tourism policy example in the main text) or far-reaching geopolitical consequences (the non-proliferation example). It is crucial at this point to note that motivations behind actions form an important determinant of legal punishment (Hart, 1958; Foucault, 1977), as shows for example in the legal distinction between manslaughter and murder. More generally speaking, it is well recognized in the fields of ethics and moral psychology that 'moral punishment', e.g. in terms of contempt or indignation, refers to motivations underlying moral actions rather than the actions themselves[9] (e.g. Alfano, 2016). This provides a clear incentive for the agent to create "reasonable doubt" (in a legal setting) or "moral wriggle room" (in a moral dilemma). Especially when the agent is uncertain about his own goals, or when he has no strong goal importance hierarchy, he may wish to avoid onlookers pinning him down on a particular goal. In all such cases, hiding your goals may be a better strategy than letting them echo through clearly in your choices.

In such situations where agents are reluctant to be transparent about their latent goals, there are various reasons why an agent would be tempted to obfuscate rather than deceive. First, the agent may simply be unaware of what are considered – by the onlooker – to be 'good' and 'bad' goals or motivations, making deceit an irrational strategy. This reason becomes even more salient when there are multiple onlookers with conflicting ideas about what is right or wrong. Second, the agent may believe that deceit is more easy for an onlooker to spot than obfuscation,

[8] A related situation to a moral dilemma concerns biases: human decision-makers are known to have several biases regarding for example gender and ethnicity (Greenwald et al., 1998). When an agent knows that his actions are being observed by an onlooker, and when he is aware that his behaviors may be biased in certain ways, he may wish to choose actions which, while still being more or less in line with the his 'biased' preferences, are difficult for the onlooker to interpret as clear evidence of biased decision-making. Such an attempt to avoid being caught out as biased is subtly different from deceit, which would involve an active attempt by the agent to signal that he is unbiased or perhaps even to signal he is 'biased' in favor of certain minorities. See Beyer and Liebe (2015) for empirical evidence of such behavior: the authors find that when an individual believes that there is an anti-Semitic consensus, he or she is more likely to be open about his or her own anti-Semitic views (if any); the absence of perceived consensus makes the individual more likely to hide such views. See Schilke and Rossman (2018) for a study into the role of obfuscation in morally sensitive choice situations.

[9] For example, when someone is being pushed over while crossing the street, that person is most likely not going to be angry with the 'aggressor', when the latter explains that his motivation for his act was to save the individual from being hurt by an oncoming car (irrespective of whether the car was actually likely to have hit the individual). Things are very different of course, when the aggressor makes it clear that his act was motivated by a wish to hurt the individual.

and more costly (to the agent) in terms of punishment than obfuscation, in case the onlooker finds out. Thirdly, deceit may in fact harm the agent by hampering his abilities to reach his goals, as was illustrated in the flirtation and non-proliferation cases.

Appendix B. Identification of the obfuscation model—aMonte Carlo analysis

Whereas the main text of the paper formalized obfuscation behavior of an individual decision-maker, in this Appendix, we move to the perspective of parameter identification by a decision analyst in the context of a dataset containing choices resulting from (possible) obfuscation-based choice behavior by a set of decision-makers. The situation we consider is one where decision-makers, onlookers and decision analyst have the following behaviors:

- The **decision-maker** makes a choice from a set of three alternatives j that are defined in terms of their scores x on two attributes; he may be concerned with obfuscation and/or with preference-aligned behavior. More specifically, the decision-maker maximizes random utility, and his utility function for alternative j is specified as $U_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \gamma \cdot I_j \left\{ H_j \geq H_i \forall i \in C \right\} + \varepsilon_j$, where $\beta_1 = 1$ and $\beta_2 = 2$. That is, an alternative's utility consists of the sum of (i) a weighted summation of the alternative's scores on the two attributes and their corresponding attribute weights, the second attribute being twice as important to the decision-maker as the first one; (ii) an indicator function which returns one if the alternative is the maximum entropy (i.e., obfuscation) option in the choice set and zero otherwise,[10] multiplied by an obfuscation weight γ; (iii) an iid Extreme Value Type I error term with variance equaling $\pi^2/6$. Note that if the obfuscation weight equals zero, the model collapses to a standard linear additive random utility maximization based Logit model. If it is positive, its size determines the extent to which obfuscation of his preferences is important to the decision-maker, compared to maximizing the preference-aligned part of utility.
- The **onlooker**[11] may be a real person or a mere mental representation in the mind of the decision-maker (think of the 'moral persona' invoked in Adam Smith's writings). Note that in case the onlooker is real, it is not her actual behavior that is of interest, but rather the decision-maker's *beliefs* regarding her behavior. His beliefs are as follows, and fall in the cognitive hierarchy constellation Level 1–Level 0: the onlooker inspects the choice made by him, and she attempts to infer, from that choice, his attribute weights β_1, β_2. She does so using the Bayesian learning scheme presented in Eqs. (1) and (2) presented in Section 2. For ease of exposition, we adopt the same settings as in the example presented in that Section. That is, the choice set contains three alternatives, there are two attribute weights, and the onlooker is uncertain about which element of the set $\{0, 1, 2\}$ represents the decision-maker's weight for any particular attribute. (Note that we tested several variations of these attribute weights, leading to similar results.) Before observing the choice, the onlooker assigns an uninformative prior probability of 1/9 to each of the following nine states of the world: $\{\beta_1 = 0, \beta_2 = 0\}$; $\{\beta_1 = 0, \beta_2 = 1\}$; $\{\beta_1 = 0, \beta_2 = 2\}$; $\{\beta_1 = 1, \beta_2 = 0\}$; $\{\beta_1 = 1, \beta_2 = 1\}$; $\{\beta_1 = 1, \beta_2 = 2\}$; $\{\beta_1 = 2, \beta_2 = 0\}$; $\{\beta_1 = 2, \beta_2 = 1\}$; $\{\beta_1 = 2, \beta_2 = 2\}$.
- The **decision analyst** receives a dataset containing 10,000 choice observations, consisting of one choice each made by 10,000 decision-makers (note that we checked that our conclusions also hold for considerably smaller datasets, e.g. containing 500 cases). Each decision-maker has the same attribute weights (i.e., $\beta_1 = 1$ and $\beta_2 = 2$ as mentioned above) but is confronted with a different choice task: attribute values x_{j1} and x_{j2} (for $j \in \{1, 2, 3\}$) were randomly – across alternatives and choice tasks – drawn from the interval [0,1]. Throughout our Monte Carlo analyses, we systematically vary the obfuscation parameter γ but keep it constant across decision-makers. The analyst identifies parameters by means of maximum likelihood estimation. We distinguish between three cases: first, the analyst may be 'naive' and believe that the decision-maker's utility function is characterized as $U_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \varepsilon_j$. That is, the analyst does not consider that the decision-maker might have been trying to obfuscate an onlooker. Second, the analyst may be 'prepared', allowing for the possibility that the decision-maker might have been trying to obfuscate an onlooker, while not knowing if and to what extent this is the case. In this case, the analyst assumes the utility function which was described further above: $U_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \gamma \cdot I_j \left\{ H_j \geq H_i \forall i \in C \right\} + \varepsilon_j$. Here, the analyst attempts to estimate attribute weights and the obfuscation parameter jointly. Third, the analyst may be 'informed' and actually know the decision-makers' obfuscation parameter γ; given perfect knowledge about this parameter, the analyst sets out to estimate the decision-makers' attribute weights. Note that this third case is rather unrealistic, and will only serve as a reference or benchmark for the other two cases.

The main question that our Monte Carlo experiment attempts to answer can be put as follows: in case $\gamma > 0$, i.e., when decision-makers have attempted to obfuscate their attribute weights from a real or imagined onlooker, would the analyst still be able to identify the obfuscation parameter, which gives the degree of obfuscation, jointly with β_1 and β_2, which give the true attribute weights which the decision-makers have attempted to hide from the onlooker?

Before presenting our results, one important remark needs to be made: entropy H_i is a function of the decision-maker's beliefs regarding uncertainty in the mind of the onlooker. As such, entropy is based on the decision-maker's beliefs as to how the onlooker will use an observed choice to update a prior distribution regarding his preferences (attribute weights) into a posterior distribution. Crucially, from the analyst's viewpoint, this entropy is a data point which may be computed based on the choice task, *before* the process of model estimation; it is *not* a function of the analyst's estimates of the attribute weights. In other words, in the process of model estimation, i.e., the process of finding the maximum likelihood attribute weights (and entropy parameter), the entropy itself is invariant; see also the probability statement in Appendix B, where this is elaborated. It should also be noted that our analyses presuppose that the analyst is aware

[10] We obtained similar outcomes based on model specifications that include Entropy directly (e.g. using Eq. (5b)), as opposed to through an indicator function. It is important to consider, in such a specification, that the variation in entropy across alternatives should preferably be roughly similar to the variation across alternatives in their attribute values. This can be achieved by, for example, re-scaling entropy differences which tend to be small compared to differences in attribute values.

[11] We distinguish between an onlooker and a decision analyst, and assume that the decision-maker is aware of the onlooker but is unaware of (or ignores) the analyst in his considerations and decision-making. Alternatively, one could study the situation where the analyst *is* the onlooker, and where the decision-maker attempts to hide his preferences from the onlooker–analyst. This alternative framing would lead to a slightly different conceptualization of the notion of obfuscation, but otherwise the results and conclusions we draw would also apply to that case.

of the decision-maker's beliefs about the onlooker's priors and about how the onlooker would update those based on the choice made by the decision-maker. This fairly restricted assumption should be relaxed in future research to explore identifiability of the obfuscation model under more lenient conditions, e.g. using continuous distributions for the attribute-weights, specified over a larger domain of possible values.

We use the newly developed R-package Apollo (Hess and Palma, 2019a,b) for our analyses; our code is downloadable at: https://github.com/szepteodora/obfuscation_identification. As a starting point for our analyses, we confirm – but do not report, for reasons of space limitations – the obvious intuition that if the analyst is naive and if the decision-makers' $\gamma = 0$ (i.e., they do not obfuscate), the true attribute weights are recovered without any problem. We then confirm – but do not report, again for reasons of space limitations – another obvious intuition: if the analyst is 'naive' and if the decision-makers' $\gamma > 0$ (i.e., they do obfuscate), the estimates for the attribute weights become biased, and increasingly so as γ gets bigger. This finding is to be expected, as in this case there is a mismatch between the utility function used by the decision-makers and the one assumed by the analyst. The straightforward and intuitive implication of this result, is that when decision-makers obfuscate and the analyst is unaware of that – and does not allow for it in the estimated choice model – estimation results will be biased. In the – admittedly unrealistic – case where the decision-makers obfuscate and the analyst is 'informed', i.e. knows the decision-makers' obfuscation parameter γ, we find (but do not report) that the true attribute weights are being recovered.

Finally, we move to the most relevant and generic case, where decision-makers obfuscate and the analyst is 'prepared', that is, he allows for the possibility that decision-makers obfuscate, but does not know whether or not and to what extent this has actually happened (Fig. B.1a and b). Fig. B.1a shows that when this is the case, the true attribute weights and the obfuscation parameter are jointly being recovered by the analyst without noticeable bias, even when the obfuscation parameter is large.

In other words, from the choices made by obfuscating decision-makers, the prepared analyst can infer the presence and degree of obfuscation, as well as the true attribute weights which the decision-makers attempted to hide from the onlooker. Fig. B.1b shows the standard errors of the estimates of the attribute weights (and of the obfuscation parameter): these again increase as a function of the size of obfuscation parameter γ. This confirms the intuitive notion that a prepared analyst can spot obfuscation behavior and simultaneously to recover the true attribute weights of an obfuscating decision-maker, but with an increasing lack of precision as obfuscation becomes more pervasive.

It needs to be emphasized here, that the analyses and conclusions presented in this Appendix are to be interpreted with care: although they show that in principle, obfuscation behavior of decision-makers need not prohibit the choice modeler from estimating his models without bias, further work is needed to show that this interpretation indeed holds in general, as opposed to only in the context of the carefully constructed Monte Carlo simulation exercise on synthetic data which was presented here.

At this point, we briefly discuss two common misunderstandings that may easily arise when inspecting the entropy-based model of obfuscation-based choice behavior that has been presented above. First, one may be tempted to believe that randomly picking an alternative without considering the attributes of the alternatives in the choice set, would be a good obfuscation strategy (especially in a repeated choice context), as it would maximize entropy in the eyes of an onlooker. Second, one may be tempted to believe that a process of variety seeking (Saviotti, 1988; Kahn, 1995; Alexander, 1997), in which alternatives are chosen which have not been chosen before – or, whose attribute values have not been chosen before – would lead to choice behavior equivalent to obfuscation maximization, in a repeated choice context. Both misconceptions are based on a single underlying misunderstanding: in the obfuscation model presented in this paper, entropy refers to the posterior probability distribution of the underlying preferences (betas, attribute weights) of the decision-maker, which he aims to hide from an onlooker. In the 'random choice' and 'variety seeking' models, the entropy refers to the *choice probabilities for alternatives*. Crucially, an obfuscating decision-maker is not so much concerned with the onlooker's knowledge about which alternative is likely to be chosen, but with her ability to understand *why* he chose a particular alternative. In this light, it is clearly the betas' entropy that counts, rather than the entropy at the level of choice probabilities; see also Eq. (3) where this notion is formalized. In fact, making random choices – as well as, albeit in a more subtle way, variety seeking – can be shown to be a poor obfuscation strategy, as it would make a choice modeler conclude that all attribute weights are zero, with high certainty (small standard errors); this would boil down to deception, not obfuscation. Obfuscation behavior as presented in this paper leads to very different choice behavior compared to either random choice behavior or variety seeking behavior.

Note that the synthetic data set that is used for the identification analyses is based on simulated choice probabilities for three alternatives, for each of the 10,000 decision-makers. (note that the term simulation in this context refers to the fact that the data are synthetic) In our synthetic dataset, the alternatives' attribute values vary across decision-makers, but each decision-maker is assumed to have the same preferences and obfuscation-related beliefs. We here present a formulation for the simulated probability that a particular decision-maker, faced with a choice set, chooses a particular alternative (hence our notation omits a subscript for decision-makers). For ease of communication, our notation differs slightly from the one used directly above: we now use the symbol β for a parameter which will be estimated by the analyst; and we use the symbol $\tilde{\beta}$ for a parameter which indirectly – i.e., through the entropy which the decision-maker believes exists in the mind of the onlooker – determines the behavior of the decision-maker, but which will not be estimated by the analyst. Another small addition in notation concerns our use of s to denote a state of the world. The decision-maker believes that the onlooker assigns a prior probability of 1/9 to each of the following nine states of the world:

$$\tilde{\boldsymbol{\beta}} = \begin{cases} \tilde{\beta}^1{}_1 = 0 & \tilde{\beta}^1{}_2 = 0 \\ \tilde{\beta}^2{}_1 = 1 & \tilde{\beta}^2{}_2 = 0 \\ \tilde{\beta}^3{}_1 = 2 & \tilde{\beta}^3{}_2 = 0 \\ \tilde{\beta}^4{}_1 = 0 & \tilde{\beta}^4{}_2 = 1 \\ \tilde{\beta}^5{}_1 = 1 & \tilde{\beta}^5{}_2 = 1 \\ \tilde{\beta}^6{}_1 = 2 & \tilde{\beta}^6{}_2 = 1 \\ \tilde{\beta}^7{}_1 = 0 & \tilde{\beta}^7{}_2 = 2 \\ \tilde{\beta}^8{}_1 = 1 & \tilde{\beta}^8{}_2 = 2 \\ \tilde{\beta}^9{}_1 = 2 & \tilde{\beta}^9{}_2 = 2 \end{cases} \tag{B.1}$$

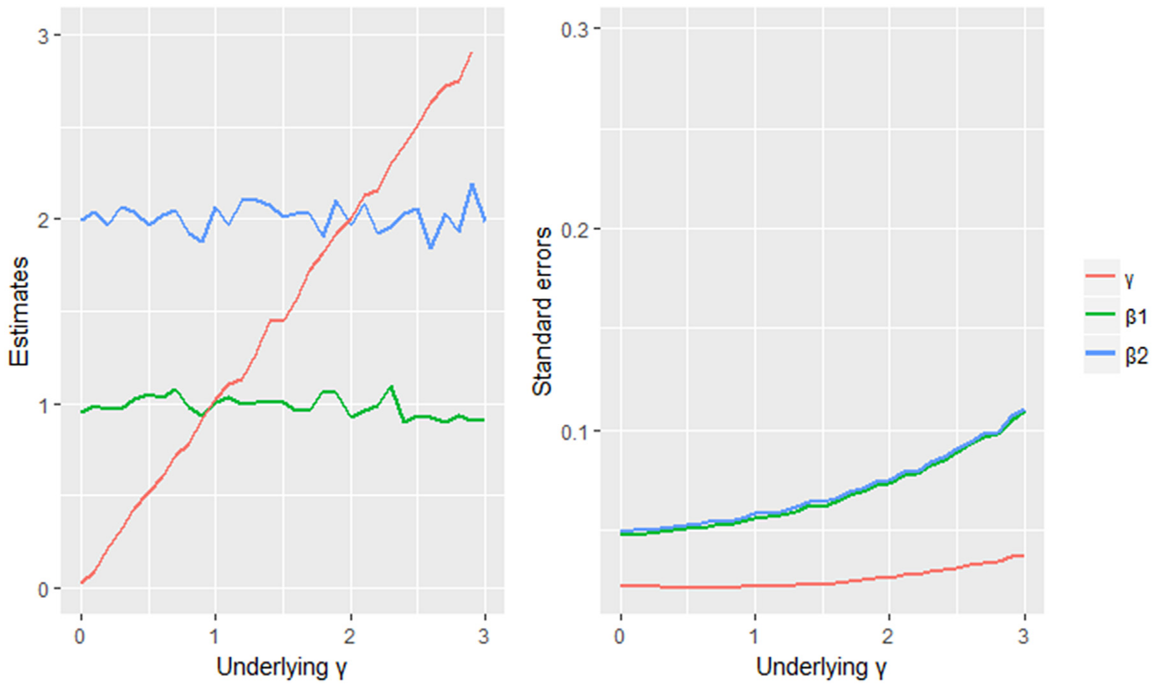


Fig. B.1. a (left hand side) and b (right hand side): estimates and standard errors in the case of a prepared analyst.

$$P\left(\tilde{\beta}^s\right)=\frac{1}{9} \qquad \forall s \tag{B.2}$$

The decision-maker believes that the onlooker assigns the following choice probability to alternative A from a set of three alternatives {A,B,C}, given a particular state of world ($\tilde{\beta}^s$) and given the attribute scores (which are also observed by the onlooker):

$$P\left(A|\tilde{\beta}^s\right)=\frac{\exp(\tilde{\beta}^s{}_1 x_{1A}+\tilde{\beta}^s{}_2 x_{2A})}{\sum_{l\in\{A,B,C\}}\exp(\tilde{\beta}^s{}_1 x_{1l}+\tilde{\beta}^s{}_2 x_{2l})} \tag{B.3}$$

This implies, that the decision-maker believes that the onlooker believes that decision-maker maximizes utility and does not obfuscate. The decision-maker also believes that upon seeing his choice for (e.g.) alternative A, the onlooker will update her prior probabilities $P\left(\tilde{\beta}^s\right)$ as to which state of the world prevails into posterior probabilities $P\left(\tilde{\beta}^s|A\right)$ using Bayes' formula:

$$P\left(\tilde{\beta}^s|A\right)=\frac{P\left(A|\tilde{\beta}^s\right)P\left(\tilde{\beta}^s\right)}{\sum_{k\in\{1,\ldots,9\}}P\left(A|\tilde{\beta}^k\right)P\left(\tilde{\beta}^k\right)} \tag{B.4}$$

Here, $P\left(A|\tilde{\beta}^s\right)$ is as given in (B.3), and $P\left(\tilde{\beta}^s\right)$ is as given in (B.2). Given these beliefs held by the decision-maker, his belief concerning the entropy in the mind of the onlooker, after she has observed his choice for alternative A equals:

$$H_A=-\sum_{s\in\{1,\ldots,9\}}P\left(\tilde{\beta}^s|A\right)\log P\left(\tilde{\beta}^s|A\right) \tag{B.5}$$

The decision-maker's choice behavior (e.g. the probability that he chooses alternative A from a set of {A,B,C}) is governed by the following Logit formula, which includes goal directed utility as well an entropy related term:

$$P_A=\frac{\exp\left[\beta_1 x_{1A}+\beta_2 x_{2A}+\gamma\cdot I_A\left\{H_A\geq H_i\forall i\in\{A,B,C\}\right\}\right]}{\sum_{i\in\{A,B,C\}}\exp\left[\beta_1 x_{1i}+\beta_2 x_{2i}+\gamma\cdot I_i\left\{H_i\geq H_j\forall j\in\{A,B,C\}\right\}\right]} \tag{B.6}$$

Here, entropy terms H are computed as given in (B.5). Similarly, choice probabilities for alternatives B and C are obtained. Based on these choice probabilities, choices are simulated for 10,000 virtual decision-makers, each making one choice given particular attribute values for all three alternatives (see settings discussed in Section 3). This data set containing 10,000 choices is then used by the analyst for model estimation. It is important to repeat there, that only parameters β_1, β_2, γ are being estimated by the analyst in the stage of model estimation. In contrast, $\tilde{\beta}^s{}_1$ and $\tilde{\beta}^s{}_2$ which are embedded in the Entropy terms (through Eqs. (B.1)–(B.5)) are pre-defined (see (B.1)), and they are not estimated. In other words, the entropy term in (B.6) is computed prior to estimation, based on each observation's attribute levels, and subsequently used as fixed input (i.e., 'data') in the stage of model estimation.

Appendix C. Instructions for the obfuscation game

I am going to explain the game. You have a sheet in front of you which has the same explanation, so you can read with me if you like to. You are about to play a game in duos. The game consists of 10 rounds. In the game, there are two roles: Decision-maker and Observer. We will randomly allocate you to be either a Decision-maker or an Observer in the first round of the game. These roles are switched between each round. So if you are a Decision-maker in the first round, you will be an Observer in the second round. So you are 5 times a Decision-maker, and 5 times an Observer. In the game you can earn real money. By participating, you get at least €15 euros. Based on how you play the game, this amount can increase. The game will consist of 10 rounds, and in each round you can earn a certain amount of money. At the end of the game, 2 rounds will be randomly drawn: one in which you were a decision-maker, and one in which you were an Observer. The money you earned in those rounds, will be paid on top of the €15. In this way, the money you earn can vary from €15–€35. The money will be transferred to your bank account after the experiment. So, as said before, you are going to play in duos, 10 rounds in total. The task in each game round is as follows. You see a matrix that displays 5 rules (the rows), and 5 actions (the columns). The cells indicate whether an action is obliged, permitted or prohibited under a certain rule. The task of the Decision-maker is as follows:

You receive a card which states what you rule is in that round. Your task is then to choose an action from the matrix. You can choose any action consistent with your rule, i.e. not forbidden by your rule. Your goal is to take an action such that the Observer remains clueless as to which rule you follow. If you choose an action that leaves the Observer clueless enough to refrain from guessing your rule, you will receive €10. If the Observer decides to guess your rule, you receive €0, irrespective of whether the Observer guesses your rule right or wrong. In sum: your aim is to make sure the Observer remains clueless as to which rule you follow, and therefore not dares to guess your rule.

The task of the Observer is as follows:

You observe the action taken by the Decision-maker and based on that information, decide whether or not to make a guess what their rule is. If you decide to guess and you guess correctly, you will receive €10. If you are clueless as to what their rule is, then you can refrain from guessing and receive €5. If you decide to guess the rule, and you guess wrong, you receive €0.

It is very important to understand as a Decision-maker that misleading the Observer is pointless. Your goal as a Decision-maker is simply to choose an action, which makes that the Observer does not dare to guess what your rule is. It makes no sense, in terms of your chance to earn money, to let him or her think that you have a certain rule, which you do not have in reality. In other words, you will earn no money for making the Observer guess wrongly; you only earn money by making him or her not guess at all. Also, there is no point in trying to win from your opponent. Your goal is to maximize the amount of money you win. There is no point in trying to do better than your opponent.

Decision-maker		Observer	
• Observer guesses your rule:	€0	• Refrain from guessing:	€5
• Observer refrains from guessing:	€10	• Guess rule of Decision-maker wrong:	€0
		• Guess rule of Decision-maker right:	€10

Here is a simplified example of the task that you do in each round:

Choice task A			
Rules	Actions		
	A1	A2	A3
R1	!	X	X
R2	✓	✓	X
R3	X	✓	✓
R4	X	!	X

Rule/actionmatrix explanation

Column: action

Row: rule

X = action is prohibited

= action is permitted

! = action is obliged

Choice task A; Rule to follow: R3

In this example, the Decision-maker is instructed to follow rule R3. This means that he cannot choose action A1, as A1 is prohibited under rule R3 (indicated by the cross). He can choose between action A2 and A3, as indicated by the checkmarks. The goal of the Decision-maker is then to choose the action that leaves the Observer most clueless as to what his rule is, so that the Observer will refrain from guessing it. If the Decision-maker had to follow R1 instead of R3 here, the ! indicates that the Decision-maker is obliged to choose A1. Remember that in the game, it is also possible that as a Decision-maker, you receive a rule which obliges you to choose a particular action (indicated by a !). This was an example of the task that you perform together in each round of the game. For the succeeding of this experiment it is important that you only say what is necessary during the game. Apart from that, we ask you to not talk to your opponent. You can also not ask questions to the facilitators during the game. It is now time to start playing the game. You have the following documents laying in front of you for this:

- A pile of choice task cards, with matrixes like in the example above
- A pile of decision rule cards, with the rules the Decision-maker has to follow
- A game form, on which you have to fill in some data after each round of the game
- A sheet with game steps, that explains the steps that you need to take in each round of the game

Please start the game by going through the sheet with game steps. Remember to end each round by filling in the game form. Once you are done with the entire game, you are free to so other things—but please do not make too much noise such as not to disturb those who are not yet done.

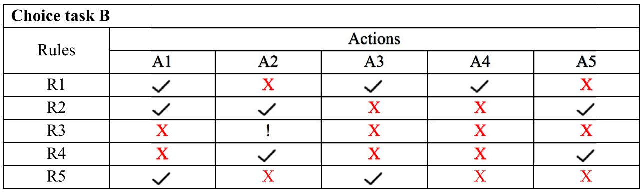
Procedure after the experiment, read out to participants at the end of the game

All participants have now completed the rounds of the game. Make sure that the entire form has been completed. As explained at the beginning, you can earn money with this experiment. This amount will vary from €15–€35. You each played 10 rounds. We will now randomly draw two rounds from these 10. We will draw one round in which you were a Decision-maker, and one round in which you were an Observer. The money you earned in those rounds will be paid on top of the €15 you received for participating. The money will be transferred to your bank account within two weeks after the experiment. We will publicly draw the numbers of the rounds that will be paid, so that you can see that this happens fairly. Before we are going to make the draws, we will collect the forms that you filled out during the game. If you want to, you can now make a picture of your game form, so that after we drew the numbers of the rounds, you can immediately see how much money you earned. So, we will now randomly draw the numbers. First we draw one number out of the possible numbers 1, 3, 5, 7, 9. Then we draw one number out of the possible numbers 2, 4, 6, 8, 10. This ensures that one round is drawn in which you were a decision-maker and one round is drawn in which you were an observer. We do this with Excel. *Show excel file*. As you can see, this file draws a random number from the possible numbers 1, 3, 5, 7, 9 and a random number from the possible numbers 2, 4, 6, 8, 10. This will be done in the form of a classical raffle. Okay, we are done now. Based on these outcomes, we will later today determine for each player how much money he or she earned and we will transfer that amount to your bank account within 2 weeks. Thank you all very much for participating in this experiment. If you have questions, you can come to me or one of the other supervisors available to ask them.

Appendix D. Choice cards for the obfuscation game

Choice task A					
Rules	Actions				
	A1	A2	A3	A4	A5
R1	✓	X	✓	X	✓
R2	X	!	X	X	X
R3	✓	X	✓	✓	X
R4	✓	✓	X	X	✓
R5	✓	✓	✓	X	✓

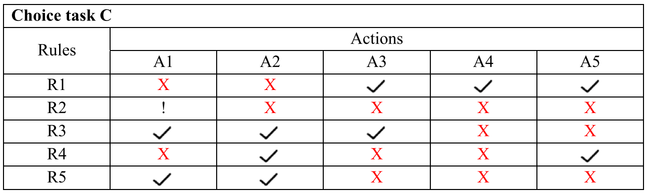
Choice task A; Rule to follow: R1



Choice task B

Rules	Actions				
	A1	A2	A3	A4	A5
R1	✓	X	✓	✓	X
R2	✓	✓	X	X	✓
R3	X	!	X	X	X
R4	X	✓	X	X	✓
R5	✓	X	✓	X	X

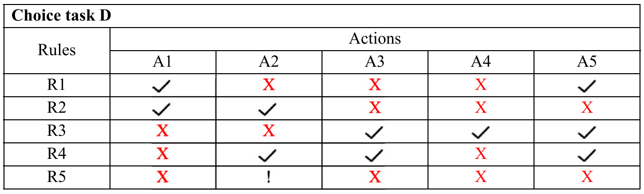
Choice task B; Rule to follow: R2



Choice task C

Rules	Actions				
	A1	A2	A3	A4	A5
R1	X	X	✓	✓	✓
R2	!	X	X	X	X
R3	✓	✓	✓	X	X
R4	X	✓	X	X	✓
R5	✓	✓	X	X	X

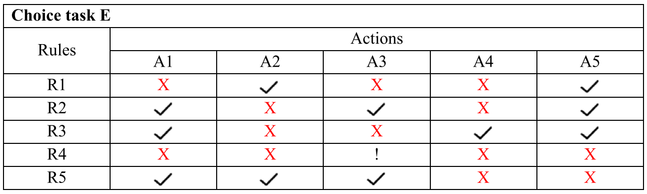
Choice task C; Rule to follow: R3



Choice task D

Rules	Actions				
	A1	A2	A3	A4	A5
R1	✓	X	X	X	✓
R2	✓	✓	X	X	X
R3	X	X	✓	✓	✓
R4	X	✓	✓	X	✓
R5	X	!	X	X	X

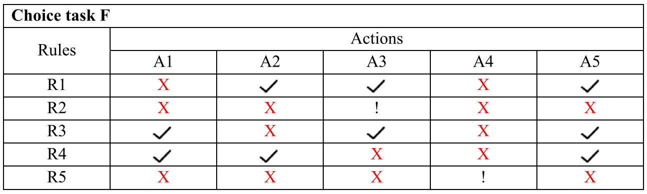
Choice task D; Rule to follow: R4



Choice task E

Rules	Actions				
	A1	A2	A3	A4	A5
R1	X	✓	X	X	✓
R2	✓	X	✓	X	✓
R3	✓	X	X	✓	✓
R4	X	X	!	X	X
R5	✓	✓	✓	X	X

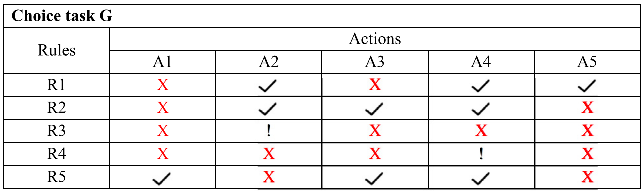
Choice task E; Rule to follow: R5



Choice task F

Rules	Actions				
	A1	A2	A3	A4	A5
R1	X	✓	✓	X	✓
R2	X	X	!	X	X
R3	✓	X	✓	X	✓
R4	✓	✓	X	X	✓
R5	X	X	X	!	X

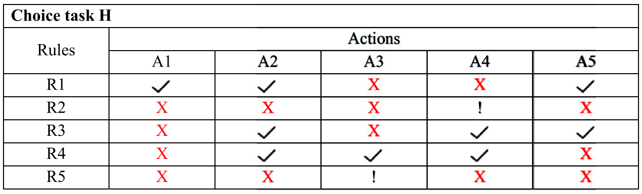
Choice task F; Rule to follow: R1



Choice task G

Rules	Actions				
	A1	A2	A3	A4	A5
R1	X	✓	X	✓	✓
R2	X	✓	✓	✓	X
R3	X	!	X	X	X
R4	X	X	X	!	X
R5	✓	X	✓	✓	X

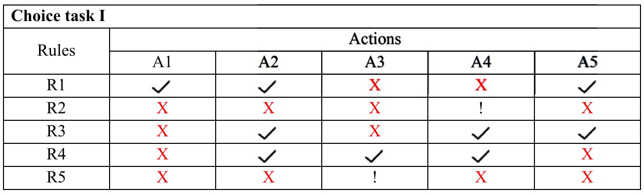
Choice task G; Rule to follow: R2



Choice task H

Rules	Actions				
	A1	A2	A3	A4	A5
R1	✓	✓	X	X	✓
R2	X	X	X	!	X
R3	X	✓	X	✓	✓
R4	X	✓	✓	✓	X
R5	X	X	!	X	X

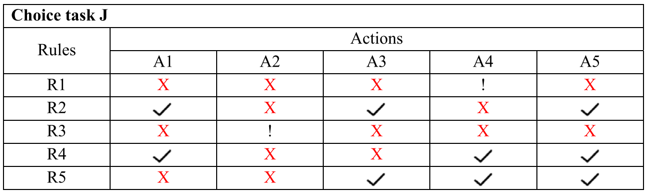
Choice task H; Rule to follow: R3



Choice task I

Rules	Actions				
	A1	A2	A3	A4	A5
R1	✓	✓	X	X	✓
R2	X	X	X	!	X
R3	X	✓	X	✓	✓
R4	X	✓	✓	✓	X
R5	X	X	!	X	X

Choice task I; Rule to follow: R4



Choice task J

Rules	Actions				
	A1	A2	A3	A4	A5
R1	X	X	X	!	X
R2	✓	X	✓	X	✓
R3	X	!	X	X	X
R4	✓	X	X	✓	✓
R5	X	X	✓	✓	✓

Choice task J; Rule to follow: R5

References

Ajzen, I., 1991. The theory of planned behavior. Organ. Behav. Hum. Decis. Process. 50 (2), 179–211.

Ajzen, I., Fishbein, M., 1977. Attitude-behavior relations: A theoretical analysis and review of empirical research. Psychol. Bull. 84 (5), 888.

Alexander, P.J., 1997. Product variety and market structure: A new measure and a simple test. J. Econ. Behav. Organ. 32 (2), 207–214.

Alfano, M., 2016. Moral Psychology: An Introduction. Polity Press, Cambridge, UK.

Aragones, E., Neeman, Z., 2000. Strategic ambiguity in electoral competition. J. Theor. Polit. 12 (2), 183–204.

Arentze, T.A., Timmermans, H.J., 2009. A need-based model of multi-day, multi-person activity generation. Transp. Res. B 43 (2), 251–265.

Bagwell, L.S., Bernheim, B.D., 1996. Veblen effects in a theory of conspicuous consumption. Amer. Econ. Rev. 34, 9–373.

Bell, D.E., Raiffa, H., Tversky, A. (Eds.), 1988. Decision-Making: Descriptive, Normative, and Prescriptive Interactions. Cambridge university Press.

Ben-Akiva, M.E., Lerman, S.R., Lerman, S.R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand, Vol. 9. MIT press.

Ben-Akiva, M., Swait, J., 1986. The Akaike likelihood ratio index. Transp. Sci. 20 (2), 133–136.

Beyer, H., Liebe, U., 2015. Three experimental approaches to measure the social context dependence of prejudice communication and discriminatory behavior. Soc. Sci. Res. 49, 343–355.

Biziou-van-Pol, L., Haenen, J., Novaro, A., Occhipinti Liberman, A., Capraro, V., 2015. Does telling white lies signal pro-social preferences?. Judgm. Decis.-Mak. 10, 538–548.

Brunton, F., Nissenbaum, H., Echizen, I., Houmansadr, A., Cote, N., Hammond, R., …Grosser, B., 2017. Obfuscation Workshop Report. In: Obfuscation Workshop 2017.

Camerer, C.F., Ho, T.H., Chong, J.K., 2004. A cognitive hierarchy model of games. Q. J. Econ. 119 (3), 861–898.

Capraro, V., Rand, D.G., 2018. Do the Right Thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. Judgem. Decis.-Mak. 13 (1), 99–111.

Carson, R.T., Groves, T., 2007. Incentive and informational properties of preference questions. Environ. Resour. Econ. 37 (1), 181–210.

Castelfranchi, C., 2000. Artificial liars: Why computers will (necessarily) deceive us and each other. Ethics Inf. Technol. 2 (2), 113–119.

Chorus, C.G., Bierlaire, M., 2013. An empirical comparison of travel choice models that capture preferences for compromise alternatives. Transportation 40 (3), 549–562.

Danaher, J., 2020. Robot Betrayal: a guide to the ethics of robotic deception. In: Ethics and Information Technology. pp. 1–12.

Davis, D.D., Holt, C.A., 1993. Experimental Economics. Princeton university press.

Edwards, W., 1954. The theory of decision-making. Psychol. Bull. 51 (4), 380.

Einhorn, H.J., Hogarth, R.M., 1981. Behavioral decision theory: Processes of judgement and choice. Annu. Rev. Psychol. 32 (1), 53–88.

Eriksson, K., Simpson, B., 2007. Deception and price in a market with asymmetric information. Judgm. Decis.-Mak. 2 (1), 23.

Forsyth, D.R., Nye, J.L., 1990. Personal moral philosophies and moral choice. J. Res. Personal. 24 (4), 398–414.

Foucault, M., 1977. Discipline and Punishment: The Birth of the Prison. Pantheon Books.

Frank, R.H., 1996. The Political Economy of Preference Falsification: Timur Kuran's Private truths, Public Lies. J. Econ. Lit. 34 (1), 115–123.

Gelman, A., Loken, E., 2013. The Garden of Forking Paths: Why Multiple Comparisons can be a Problem, Even when there is No "Fishing Expedition" or "P-Hacking" and the Research Hypothesis was Posited Ahead of Time. Department of Statistics, Columbia University.

Georgeff, M., Pell, B., Pollack, M., Tambe, M., Wooldridge, M., 1998. The belief–desire-intention model of agency. In: International Workshop on Agent Theories, Architectures, and Languages. Springer, pp. 1–10.

Gigerenzer, G., 2010. Moral satisficing: Rethinking moral behavior as bounded rationality. Top. Cogn. Sci. 2 (3), 528–554.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L., 1998. Measuring individual differences in implicit cognition: the implicit association test. J. Personal. Soc. Psychol. 74 (6), 1464.

Haidt, J., 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. 108 (4), 814.

Hart, H.M., 1958. The aims of the criminal law. Law Contemp. Probl. 23 (3), 401–441.

Hess, S., Palma, D., 2019a. Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. J. Choice Model..

Hess, S., Palma, D., 2019b. Apollo version 0.0.8, user manual. www.ApolloChoiceModelling.com.

Houthakker, H.S., 1950. Revealed preference and the utility function. Economica 17 (66), 159–174.

Hurwicz, L., 1973. The design of mechanisms for resource allocation. Amer. Econ. Rev. 63 (2), 1–30.

Jarzabkowski, P., Sillince, J.A., Shaw, D., 2010. Strategic ambiguity as a rhetorical resource for enabling multiple interests. Hum. Relat. 63 (2), 219–248.

Jolink, A., Niesten, E., 2020. Credibly reducing information asymmetry: Signaling on economic or environmental value by environmental alliances. In: Long Range Planning. 101996.

Kagel, J.H., Roth, A.E. (Eds.), 2016. The Handbook of Experimental Economics, Volume 2: The Handbook of Experimental Economics. Princeton university press.

Kahn, B.E., 1995. Consumer variety-seeking among goods and services: An integrative review. J. Retail. Consum. Serv. 2 (3), 139–148.

Keeney, R.L., Raiffa, H., 1993. Decisions with Multiple Objectives: Preferences and Value Trade-Offs. Cambridge university press.

Kivetz, R., Netzer, O., Srinivasan, V., 2004. Alternative models for capturing the compromise effect. J. Mark. Res. 41 (3), 237–257.

Kono, D.Y., 2006. Optimal obfuscation: Democracy and trade policy transparency. Amer. Polit. Sci. Rev. 100 (3), 369–384.

Kuran, T., 1997. Private Truths, Public Lies: The Social Consequences of Preference Falsification. Harvard University Press.

Loewenstein, G., 1999. Experimental economics from the vantage-point of behavioral economics. Econ. J. 109 (453), 25–34.

Luce, D., 1959. Individual Choice Behavior: A Theoretical Analysis. Wiley.

Marley, A.A.J., Swait, J., 2017. Goal-based models for discrete choice analysis. Transp. Res. B 101, 72–88.

McConnell, K.E., 1995. Consumer surplus from discrete choice models. J. Environ. Econom. Manage. 29 (3), 263–270.

McFadden, D., 1974. Conditional logit analysis of qualitative choice-behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

McFadden, D., 2001. Economic choices. Amer. Econ. Rev. 91 (3), 351–378.

Nisbett, R.E., Wilson, T.D., 1977. Telling more than we can know: Verbal reports on mental processes. Psychol. Rev. 84 (3), 231.

Page, B.I., 1976. The theory of political ambiguity. Amer. Polit. Sci. Rev. 70 (3), 742–752.

Prelec, D., 2004. A Bayesian truth serum for subjective data. Science 306 (5695), 462–466.

Samuelson, P.A., 1948. Consumption theory in terms of revealed preference. Economica 15 (60), 243–253.

Sandorf, E.D., Chorus, C., van Cranenburgh, S., 2019. ObfuscatoR: Obfuscation Game Designs. R package version 0.2.0, URL https://CRAN.R-project.org/package=obfuscatoR.

Saviotti, P.P., 1988. Information, variety and entropy in technoeconomic development. Res. Policy 17 (2), 89–103.

Schilke, O., Rossman, G., 2018. It's only wrong if it's transactional: Moral perceptions of obfuscated exchange. Am. Sociol. Rev. 83 (6), 1079–1107.

Sen, A.K., 1971. Choice functions and revealed preference. Rev. Econom. Stud. 38 (3), 307–317.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.

Simonson, I., 1989. Choice based on reasons: The case of attraction and compromise effects. J. Consum. Res. 16 (2), 158–174.

Smaldino, P.E., Flamson, T.J., McElreath, R., 2018. The evolution of Covert Signaling. Sci. Rep. 8 (1), 4905.

Small, K.A., Rosen, H.S., 1981. Applied welfare economics with discrete choice models. Econometrica 10, 5–130.

Smith, V.L., 1976. Experimental economics: Induced value theory. Amer. Econ. Rev. 66 (2), 274–279.

Sunstein, C.R., 2005. Moral heuristics. Behav. Brain Sci. 28 (4), 531–541.

Swait, J., Marley, A.A., 2013. Probabilistic choice (models) as a result of balancing multiple goals. J. Math. Psych. 57 (1–2), 1–14.

Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge university press.

Tversky, A., 1972. Elimination by aspects: A theory of choice. Psychol. Rev. 79 (4), 281.

Van't Veer, A., Stel, M., van Beest, I., 2014. Limited capacity to lie: Cognitive load interferes with being dishonest. Judgm. Decis.-Mak. 9 (3), 199–206.

Von Stackelberg, H., 2010. Market Structure and Equilibrium. Springer Science & Business Media, (translated to English).

Walker, J., Ben-Akiva, M., 2002. Generalized random utility model. Math. Social Sci. 43 (3), 303–343.

Zurek, T., 2017. Goals, values, and reasoning. Expert Syst. Appl. 71, 442–456.