You can fool some of the people some of the time but you can help everyone forever: learning and effective nudging*

Mattias Forsgren[1], Gustav Karreskog Rehbinder[2], and Benjamin Mandl[3]

[1]Uppsala University, Department of Psychology
[2]Uppsala University, Department of Economics
[3]Jua.ai

July 18, 2024

Abstract

Why do nudges sometimes fail to deliver the promised behaviour change? We argue that part of the explanation may be that people learn whether the nudge is guiding them towards their goals or not. In an experiment, we show that participants quickly learn to choose in accordance with a nudge proportionally to how well it predicts the superior option. This illustrates a more general point: unless choice architects align their nudging with the goals of the nudgee, the latter's capacity to learn and make inferences may allow them to come up with strategies to avoid being nudged.

Keywords: Nudges, learning, default effect, decoy effect, experiment **JEL:** D91, D82, C91

*Corresponding author: Gustav Karreskog Rehbinder, Department of Economics, Uppsala University, Kyrkogårdsgatan 10, Ekonomikum, 753 12 Uppsala. Email: gustav.karreskog@nek.uu.se.

We thank Anna Dreber Almenberg, Ola Andersson, Magnus Johannesson, Erik Mohlin, Nurit Nobel, as well as seminar participants at the Arne Ryd workshop and the Swedish conference in economics. This work was supported by the Jan Wallander and Tom Hedelius Foundation, the Knut and Alice Wallenberg Research Foundation, the Swedish Research School of Management & IT, and the Marcus and Amalia Wallenberg Foundation.

1 Introduction

Choice architecture design is widely used to bring about a desired behaviour by changing the context of a choice problem. Two common context interventions are (i) manipulating the default alternative and (ii) adding an inferior "decoy" option to make some target option more attractive (the former being more commonly used by behavioural policy makers and the latter by marketers). Why and when such "nudges" work is not fully understood: suggestions include that they increase the salience of the target option, provide a reference point that induces loss aversion, or constitute (implicit) recommendations (Jachimowicz et al., 2019). Similarly, it is not clear why nudges fail. Recent popular science work has noted how nudging in the field often fails to deliver the promised behavioural change observed in the laboratory (Mazar and Soman, 2022).[1] Sunstein (2017) lists strong preferences and potential "counter-nudges" as main causes. Here, we argue that the explanation may also be that the decision maker learns the relation between observable characteristics ("cues") of nudges (e.g. whether an option is the default) and outcomes (whether the default option was any good). If a cue predicts beneficial outcomes (e.g. the default option being superior to the alternatives), the decision maker may come to rely on it to make their choice. If a cue does not predict beneficial outcomes (the cue is unrelated to option superiority) or even inferior outcomes (the cue is related to option inferiority), the decision maker may safely ignore it or actively avoid it. That humans have this capacity is psychologically uncontroversial, as we return to below. We demonstrate this ability experimentally in the context of nudging by randomly varying the probabilistic cue-outcome relationship (with what probability a cue predicts whether an outcome is superior to the alternatives) of the default and decoy cues and find that people's behaviour is adapted to cue predictivity: the stronger the historic cue-outcome relationship, the more often people choose according to it. We also include a cue that is not taken from the nudge literature to illustrate the generality of this phenomenon. Our most fundamental conclusion is this: nudges which run contrary to a decision maker's goals may occasionally and/or transiently affect behaviour, but may eventually be overcome by learning. Nudges which direct choice towards options aligned with the decision maker's goals may affect behaviour permanently and will be reinforced by learning. As far as (cue reliant) nudging goes, you can fool some of the people some of the time but you can help everyone forever. We immediately concede a limitation: we have tested specific nudges in a specific paradigm and there are many situations where learning is not feasible (e.g. due to lack of feedback). We do emphasise, however, that learning is just one class of of many cognitive mechanisms (Fodor, 1983) at people's disposal, each suitable for particular settings. We thus expect analogous results to

[1] A controversy surrounding a recent meta analysis of nudges (Mertens et al., 2021) has highlighted that the literature also suffers from severe publication bias (Mertens et al., 2022; Szaszi et al., 2022; Maier et al., 2022; Bakdash and Marusich, 2022). Another part of the problem may thus be that many nudges simply do not even work in the lab.

obtain in other settings too, deriving not from the same cognitive mechanism as is active here but from others.

We now describe the two nudges taken from the literature that we will include in our experiment. The Swedish pension system allows citizens to select where to invest a part of their pension savings but they are defaulted into a low-fee public fund ("AP7", https://www.ap7.se/english/). This fund has outperformed the private funds since inception. More than 5 million out of 6.4 million working age Swedes choose the default fund (Cronqvist, Thaler and Yu, 2018). Clearly, the nudge works. Our example of an ineffective default nudge comes from personal experience: when booking a Ryan Air flight there are multiple expensive add-on purchases, such as extra travel insurance or seat upgrades, selected by default. These nudges do not affect the second and third authors' purchasing behavior:[2] we de-select all the defaults. Why does the default nudge "work" in one setting and not the other? One important difference between our examples is whether the decision maker can trust the choice architect to have their best interest at heart.[3] The implicit recommendation effect is clear for both situations but Swedes would have learned to trust the recommendations from the Swedish government and the second and third authors have learned to distrust the recommendations from a profit maximizing airline.

The attraction effect exists when adding some dominated option (a decoy) to an option set increases the probability of a participant selecting the dominating option (the option that has a decoy) vis-à-vis an option that dominates neither the decoy nor the option that has a decoy. This constitutes a violation of independence from irrelevant alternatives, which states that the relative choice preference between any pair of options is unaffected by adding or removing other options in a bundle (Debreu, 1960; Morgenstern and Von Neumann, 1953). Although the effect has been demonstrated many times (e.g. Farmer et al., 2016; Müller, Schliwa and Lehmann, 2014; Castillo, 2020; Lichters et al., 2017), even with incentivised designs (e.g. Lichters et al., 2017), failures to replicate the effect (e.g. Yang and Lynn, 2014) have led to studies qualifying the conditions under which it occurs (e.g. Frederick, Lee and Baskin, 2014). In particular, it seems important to tune the differences between options (e.g. Padamwar, Dawra and Kalakbandi, 2019; Kaptein, Van Emden and Iannuzzi, 2016) such that they exist within some critical zone where the effect can be observed.

There are (at least) two prominent psychological explanations for why an attraction effect can be observed: the decoy could provide the decision maker with a convenient argument for

[2]The first author has never flown Ryan Air.

[3]While we focus on the alignment of incentives between choice architect and decision maker, there may of course be several contributing factors to why the pension default is especially effective: actively opting out of the default plan requires some effort (logging in to the pension platform and making an active fund choice). Social norms or salience of the public pension fund might also play a role. We suppose such reasons affect the strength of any nudge independently of any learning.

justifying their choice (Shafir, Simonson and Tversky, 1993; Dietrich and List, 2016; Gomez et al., 2016; Simonson, 1989) or it could arise from how decision makers weigh together the pieces of numerical information (Roe, Busemeyer and Townsend, 2001; Trueblood, Brown and Heathcote, 2014). We do not contest these explanations, but our empirical results suggest that learning can either moderate the attraction effect, in so far as it appears, or generate a similar effect.

We believe that the explanation we provide here could account for several conspicuous examples of when nudging fails: Beshears et al. (2010) show that unusually large default contribution rates lead to the majority of employees opting out and shifting to a lower contribution rate. Bronchetti et al. (2011) show that a default nudge intended to make participants invest tax refunds into saving funds fails in a sample where 79% of the decision makers had planned to use the refund for consumption. Altmann et al. (2019) show that defaults in a donation experiment do not result in larger aggregate donations. These examples, we suspect, may have in common that the goals of the choice architect did not align with the goals of the decision makers. When the nudge-outcome relation is (thought to be) known beforehand (e.g. the relation between savings and consumption), the decision maker needs no or minimal experience to adapt their choices but can make inference from their knowledge. When the predictivity of the nudge is unknown, they may learn it.

2 The psychology of probability and cue learning

A now classic review by Peterson and Beach (1967) presented people as "intuitive statisticians" who are able to learn statistics of observed samples of data. Recent work has highlighted our ability to learn (and track) parameters of non-stationary distributions as they change (e.g. Forsgren, Juslin and van den Berg, 2023; Nassar et al., 2010; McGuire et al., 2014; Norton et al., 2019). The tasks in this literature involve estimating, for example, the current mean of a distribution of outcomes. To use such estimates to make a choice we must also have some, at least ordinal, valuations of those outcomes.

One way of stylising such valuations is through the "lens model" (Brunswik, 1952, 1956). Here, decisions and judgements (e.g. what option to select) are conceived of as directed towards some "distal" property which is not directly observable at the time of choosing (e.g. the utility received from consuming some product). The nature of the distal object has to be inferred from "proximal" cues that *are* directly observable (e.g. price, some measure of quality). Cues are only imperfect predictors of properties of options and the challenge for the organism is to learn how much weight to put on each cue. The best one can do is to adapt one's weighting of cues such that it is equal to how strongly a cue actually predicts a property (Hursch, Hammond and Hursch, 1964). People appear to be decent at this but performance depends on characteristics of the task and feedback (see Karelaia

and Hogarth, 2008, for a review). When cues have a simple (e.g. linear) relation to the outcome variable (e.g. Juslin et al., 2003*a*), this learning is so-called "declarative", which means that the extracted relation or rule is accessible to consciousness and can be verbalised (see Tulving, 1972; Shayna Rosenbaum, Kim and Baker, 2017), allowing participants to extrapolate beyond the observed cue values.

Although it is of course an empirical question, the above invites us to propose that it would be psychologically unremarkable if it turned out that people (i) learn what properties that defaults, decoys and visual characteristics predict, (ii) how reliably they do so, and (iii) apply those learnings to help them reach their goals. This is what we will demonstrate below.

3 Experiment

We study the effect of predictivity of cues on decision making in a preregistered online experiment.[4] Participants are given 40 variations of a simple decision task. Participants are randomized into one of three cue types: the decoy, default, or rule condition. The decoy and default cue types are described in the Introduction. The rule cue type is included to see if any effect also obtains for a more arbitrarily selected aspect of the decision environment. If so, it would be easier to argue that the results do not arise because of some characteristics specific to the cue types from the nudge literature, and thus may generalise to additional cue types. The treatment is how predictive the decoy/default/rule is (the probability of the cue indicating the superior option), which we randomise for each participant. We view each cue type condition as a replication of the effect of our treatment on choice behaviour under a different "nudge".

3.1 Experiment design

The task is based on Crosetto and Gaudeul (2016) and Trueblood et al. (2013). Participants are asked to choose one of three geometric shapes (which may be vertical rectangles, horizontal rectangles, vertical ellipses, horizontal ellipses, squares, or circles) with different areas (e.g. 2 cm, 2.5 cm) presented on a grid. Each option comes at some price, in experimental units. Upon choosing an option, the participant receives a payoff equal to the area minus the price of the chosen option. The participant also receives full feedback (Camilleri and Newell, 2011): the areas of all shapes and their corresponding payoffs are

[4]The default and decoy cue types were preregistered at `https://doi.org/10.17605/OSF.IO/DS4XP`. Data for the rule condition was collected later and preregistered at `https://doi.org/10.17605/OSF.IO/852ZV`. A preregistered pilot experiment is available in Mandl (2022).

presented on screen after each choice. At the end of the experiment, the experimental units are converted into pound sterling (GBP) at a predetermined exchange rate of 1500 units to GBP 1. Participants must wait at least 5 seconds per trial before they can submit a choice, to invite some consideration. See Figure 1 for the task interface and feedback screen and the online appendix for task instructions.

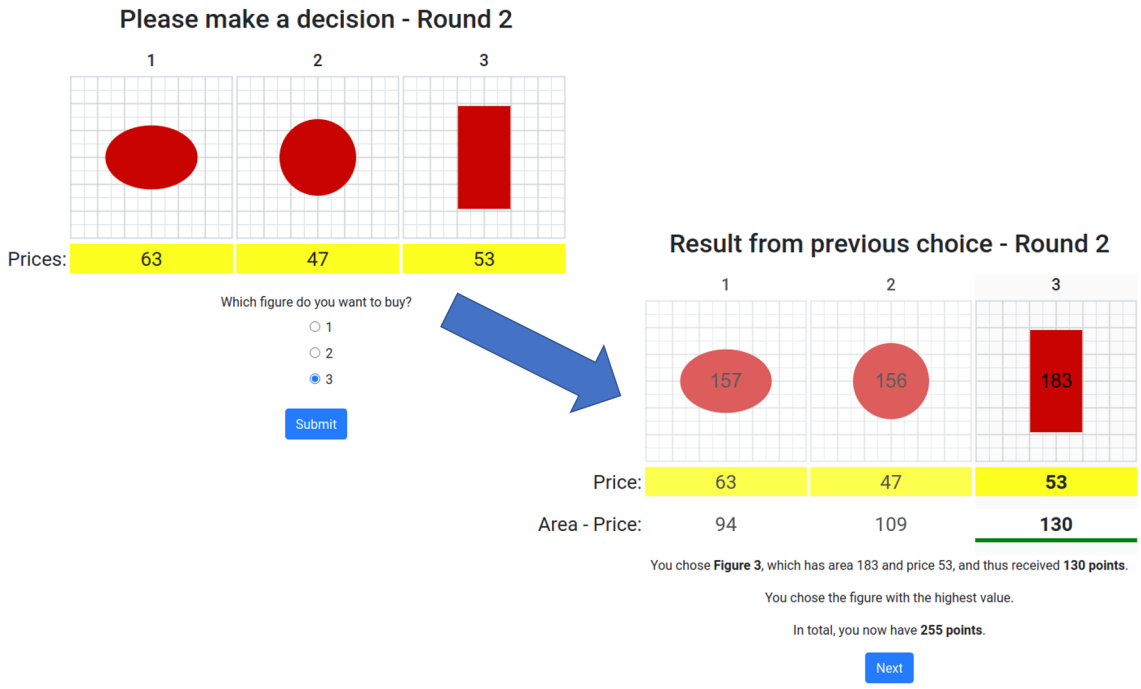

Figure 1: The task as it appeared to participants. A participant in the default condition is about to choose option 3, a rectangle of area 183 and price 53. Once they have done so it is revealed that this yields 130 points. They are also informed that they chose the superior option, how many points the alternatives would have yielded, and their total score.

We generalize the design of earlier decoy and default studies by generating each option payoff by drawing from predetermined random distributions: first, three payoffs are drawn from the same normal distribution and then ranked. Then, the price of each option is drawn from another normal distribution. Each option's area is determined by adding the payoff and the price. Lastly, we randomly select each option's geometric shape. These features exhaustively define each option. We vary the treatment "assigned predictivity" by randomising the probability of the cue indicating the superior option. In 28 of the rounds, the cue is generated according to the assigned predictivity ("cue treatment rounds"). In the remaining 12 rounds, the cue is on average unpredictive (uncorrelated with value) for all participants and identically generated for all participants with the same cue type. We

will use these 12 test rounds to test our hypotheses (to observe measurement invariance, see Meredith, 1993). Treatment and test rounds are mixed and presented in random order. In trials where the cue does not indicate the superior option, it indicates one of the other two options. The predictivity of a cue is thus never stated but must be learnt from experience.

For any given trial, we can indicate whether the cue did predict the superior option in that round using a dummy variable:

$$\text{cue predictive} = \begin{cases} 1 \text{ if superior option has a decoy/is default/follows rule} \\ 0 \text{ otherwise} \end{cases}$$

we can calculate the share of such predictive rounds observed by the participant so far. We call that variable historical predictivity.

$$\text{historical predictivity}_r = \sum_{i=1}^{r-1} \frac{\text{cue predictive}_i}{r-1}$$

where r is rounds from 1 to 43,
rounds 1 to 3 are practice rounds,
rounds 4 to 43 are the experiment rounds

This measure will be the independent variable in our analyses. We will thus not investigate the effect of the *assigned* predictivity (cf. intention to treat) but of the historical predictivity that participants actually observe.

3.1.1 Decoy condition

In the 28 treatment rounds, options are constructed as follows. Option features are randomised (see Experiment design) to generate three options. With probability p equal to the assigned predictivity, the superior option is selected to have a decoy. If so, one of the remaining options (either the second or the third best) is chosen with 50% probability to be the superior option's decoy. The decoy option is then adjusted to be dominated by the superior option and assume its geometric shape since this facilitates ranking them by area (Natenzon, 2019; Smith, 1969), tentatively ruling out that any absence of an effect would arise simply due to limitations in perceptual discrimination rather than lack of learning. The option that *has* a decoy and the option that *is* a decoy thus always have the same shape. With probability $1 - p$, the second best option is chosen to have a decoy. If so, the

worst option is the decoy. For the 12 test rounds, the options are constructed as follows. First, either the superior or the second-best option is determined to have a decoy with equal probability. That option is then assigned the worst option as its decoy. The decoy is adjusted as before. In test rounds, an option having a decoy is thus not indicative of whether it is the superior or second best option but it does indicate that it is not the worst option.

3.1.2 Default condition

In the 28 treatment rounds, options are constructed as follows. Option features are randomised (see Experiment design) to generate three options. With probability p equal to the assigned predictivity, the superior option of the three is pre-selected and participants thus only have to click the "submit" button to select it. In $1 - p$ of cases, one of the two inferior options is pre-selected. In the 12 test rounds, the options are constructed as follows. Option features are randomised to generate three options. One of the options is pre-selected by default with a uniform probability over the three options. The default is thus not indicative of whether an option is superior, second best or third best in those rounds.

3.2 Rule condition

In the 28 treatment rounds, one geometric shape per participant is randomised to be the "indicator shape" (e.g. square). Then, option features are randomised to generate three options but no option is allowed to assume the indicator shape. After this, with probability p the superior option is forced to assume the indicator shape. In $1 - p$ of cases, one of the two inferior options is forced to assume the indicator shape. The rule is thus that one of the shapes indicates the superior option with some probability p. In the 12 test rounds, shapes are randomised as per Experiment design and a randomly selected option is then forced to assume the indicator shape. The indicator shape is thus not indicative of whether an option is superior, second best or third best in those rounds.

3.3 Hypotheses

Our hypotheses establish whether the historical predictivity over all rounds predicts the propensity to follow the cue in test rounds, which are identical between participants.

Main Hypotheses - decoy/default/rule: *Participants who experience a higher historical predictivity follow the decoy/default/rule more often in the 12 test rounds.*

To test these hypotheses, we subset the data in the decoy, default, and rule condition, respectively, to obtain only the 12 test rounds. Separately for each condition, we then regress the binary variable of whether the option with a decoy/default option/option indicated by the rule was chosen on the predictivity. For each regression, we predict a positive coefficient of the predictivity measure using a two-sided t-test and a significance level α of 0.05.

3.4 Procedure

Data was collected on Prolific Academic, a popular online social science laboratory (see Palan and Schitter, 2018; Peer et al., 2017, for overviews).[5] We had a rolling recruitment with the pre-registered goal of having at least 300 participants for each cue type. We recruited a total of 1010 participants for our study. After applying our exclusion criteria, the final sample contains 302 participants in the decoy cell, 306 in the default cell, and 301 in the rule cell. The average age of these participants is 29.3 and 47% are women.

Participants gave informed consent, received identical instructions in all conditions, and were able to practice the task for three rounds. The practice rounds were generated according to the participants' respective treatment levels: their cue type and their assigned predictivity. Participants then needed to pass a comprehension quiz with 4 questions. If they failed it more than 10 times, their participation was terminated.

Participants collected payoffs (the area of their selected option minus its price) in experimental units in each of the 40 rounds. After the experiment, the points were converted into GBP according to an exchange rate known to the subjects (1500 units to GBP 1). The average payment was GBP 2.85.

3.5 Results

We find a strong and statistically significant treatment effect in our preregistered regressions for all cue types, see Table 1. An increase of historical predictivity by 10 percentage points results in an increase in the probability of choosing the option with a decoy, the default, and the option indicated by the rule by 1.29%, 1.58%, and 2.21%, respectively.

To visualize our results, we can graph the linear relationship between the mean probability of choosing the option that has a decoy/is the default/is indicated by the rule and the historical predictivity, see Figure 2.

[5]Data for the decoy and default conditions was collected from September through October of 2021. Data for the rule condition was collect through March of 2024.

Table 1

	Dependent variable:		
	Option with decoy was chosen	Default option was chosen	Option indicated by rule was chosen
	(1)	(2)	(3)
Constant	0.476	0.283	0.265
	(0.021)	(0.020)	(0.020)
Cumulative predictivity	0.129	0.158	0.221
	(0.037)	(0.039)	(0.044)

Note: Standard errors in parentheses, robust and clustered on individual level. $p < 0.001$ for all estimates.

What those relationships should look like, in the presence or absence of a learning effect, differs slightly between conditions.

For the default and rule conditions it is quite straightforward. If the participants ignore the cue, the shape that is the default/is of the indicator shape should be chosen 1/3 of the time in the test rounds. If the participants respond to the cue, the shape that is the default/is of the indicator shape should be chosen 1/3 of the time only when historical predictivity is 1/3 and increase with historical predictivity.

For the decoy effect, understanding what relationship we should obtain is slightly more intricate. First note that the option that has a decoy is never the worst option and that the option that is a decoy is never the best option. If a participant observes this, they should never choose the option that is a decoy, effectively reducing the task to a choice between two options (the option that has a decoy and the third, remaining option). Considering those two options, the option that has a decoy can only ever be the best or second best option (since the option that is a decoy is always dominated by the option that has a decoy). The remaining option however (which neither has a or is the decoy), can be the best, second best, or worst option, while the option that has a decoy can only ever be the best or second best. This means that even if the option that has a decoy is the best option only half of the time, it might still have a higher expected payoff than the remaining option. In the online appendix we show that already at a historic predictivity of 36%, the option that has a decoy has a higher expected payoff. If participants ignore the cue, the option that has a decoy should thus be chosen 1/2 of the time. If particpants respond to the cue, the option that has a decoy should be chosen 1/2 of the time only when historical

predictivity is 36% and increase with historical predictivity.

In all three conditions, we observe relationships consistent with participants responding to the cue. For the decoy condition, the regression line intersects 1/2 when historical predictivity is about 36% and for the default and rule conditions it intersects 1/3 when historical predictivity is about 1/3 (Figure 2).

4 Discussion

Our most fundamental conclusion is that the effectiveness of the nudges investigated here scaled with how useful they were for a decision maker in helping them select the superior option. With minimal training, participants learned to ignore cues that were not conducive to their goal. We suggest that this kind of learning might explain why nudging sometimes fails to deliver the desired behaviour change: the choice architect tries to nudge the decision maker in a way that is not aligned with the latter's goals. The decision maker learns this and invents strategies to avoid being nudged. To be successful, liberal paternalists must (possibly) not only avoid restricting the options of individuals. They must also promote the individuals' self-perceived goals.

Here, we have investigated a specific paradigm which affords learning from observed outcomes. There are of course many situations where such learning is not workable, for example due to a lack of feedback or the decision situation being unique or very rare. We therefore emphasise that there is a host of other mechanisms through which humans can learn (e.g. Busemeyer and Myung, 1988; DeLosh, Busemeyer and McDaniel, 1997) or make use of knowledge (e.g. Tenenbaum et al., 2011; Johnson-Laird, Khemlani and Goodwin, 2015), each evolved to fit a different niche (Marewski and Schooler, 2011). In so far as the relevant mechanism is sufficient to identify the contingency between the nudge and the self-perceived quality of the outcome, we expect to see results analogous to what we find here: if it becomes transparent to the decision maker that the nudge is not guiding them towards their goal, they become motivated to device a strategy to avoid it.

Our findings are related to a recent discussion on the use of "bad" nudges by private businesses or nudges that appear partisan to influence the consumers' or voters' behaviours (e.g. Tannenbaum, Fox and Rogers, 2017; Thaler, 2015). Examples include default cookie settings in GDPR compliant cookie banners, default newsletter subscriptions, and add-on insurance purchases. Our results invite the idea that decision makers and consumers may adapt to avoid these nudges even when stakes are small.

They also qualify the allure of nudges vis-à-vis traditional economic interventions such as taxation. Hagmann, Ho and Loewenstein (2019) suggest that introducing environmental

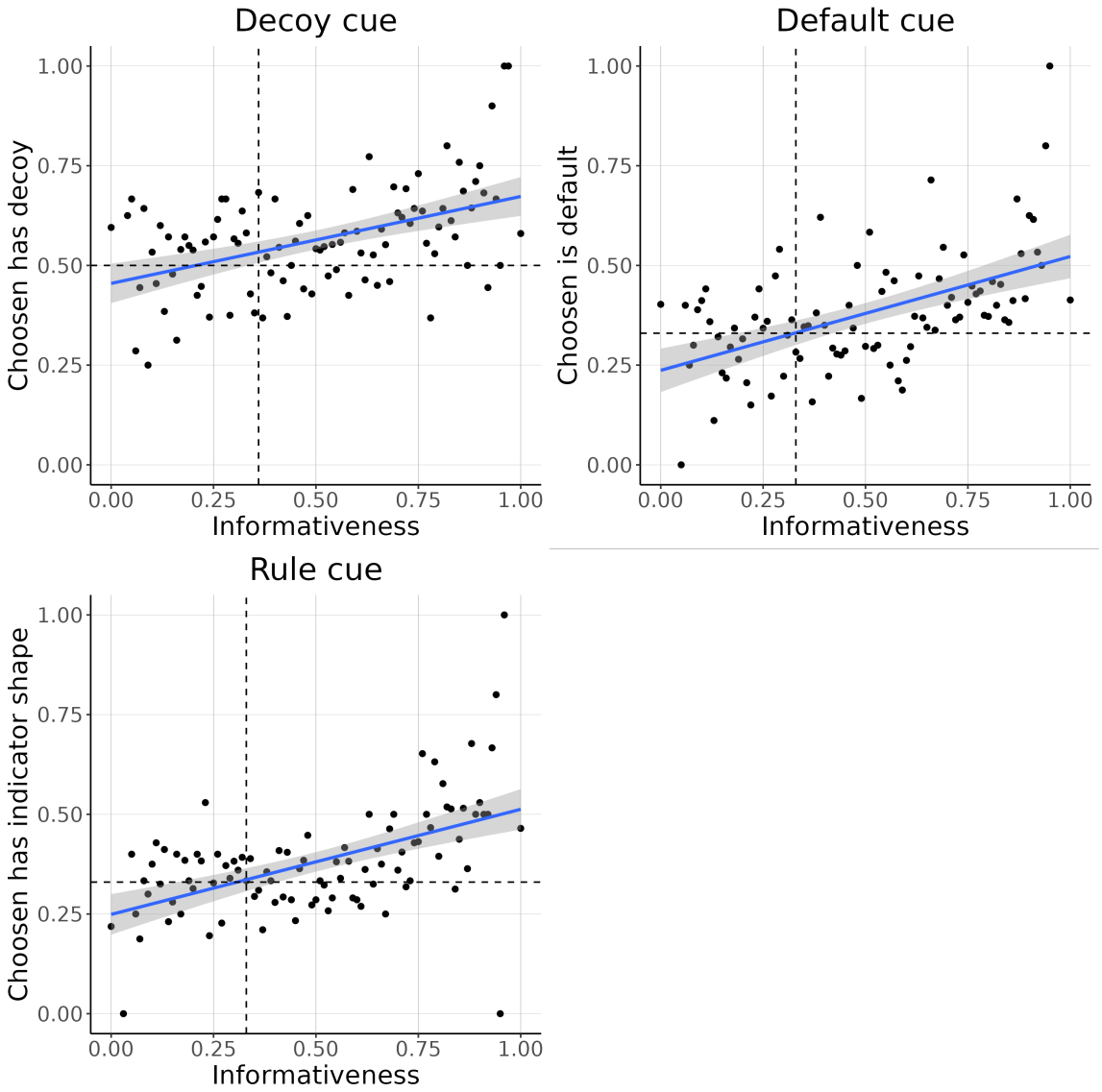

Figure 2: Regressions of historical predictivity on choosing the option that has a decoy/is the default/is of the indicator shape. The latter are binary variables, making individual observations difficult to visualise in a scatterplot since they will all sit at 0 or 1. We therefore group all observations by historical predictivity rounded to two decimal places. Each marker is thus a local average for a small subset of observations at each level of historical predictivity.

nudges may decrease support for carbon taxation. We suggest that, unless individuals are already motivated to cut carbon emissions, the effect of some nudges may decrease over time. It thus seems perilous to substitute nudges for incentivising measures like taxation, especially for urgent issues like climate change where we have little time to row back on failed policies. Nudging, we argue, should be viewed as a *complement* rather than substitute.

5 Conclusion

We have shown that participants quickly adapt their susceptibility to being nudged when they make a few repeated decisions with feedback. This is an illustration of the more general point we want to make here: humans are capable of learning from experience and using knowledge to make inferences. These capacities allow dynamic adaptation to features of our decision environment. This might make it challenging to nudge people into making decisions with outcomes that go against their self-perceived goals, since they may eventually adapt to avoid the nudge. However, it also suggests that a choice architect can sustainably and perennially nudge decision makers *towards* their self-perceived goals (cf. "boosting", Hertwig, 2017).

References

Altmann, Steffen, Armin Falk, Paul Heidhues, Rajshri Jayaraman, and Marrit Teirlinck. 2019. "Defaults and Donations: Evidence from a Field Experiment." *The Review of Economics and Statistics*, 101(5): 808–826.

Bakdash, Jonathan Z., and Laura R. Marusich. 2022. "Left-truncated effects and overestimated meta-analytic means." *Proceedings of the National Academy of Sciences*, 119(31).

Bang, H. Min, Suzanne B. Shu, and Elke U. Weber. 2020. "The Role of Perceived Effectiveness on the Acceptability of Choice Architecture." *Behavioural Public Policy*, 4(1): 50–70.

Beshears, John, James J. Choi, David Laibson, and Brigitte C. (Brigitte Condie) Madrian. 2010. "The Limitations of Defaults."

Bolton, Gary, Eugen Dimant, and Ulrich Schmidt. 2018. "When a Nudge Backfires: Using Observation with Social and Economic Incentives to Promote Pro-Social Behavior." *SSRN Electronic Journal*.

Bronchetti, Erin, Thomas Dee, David Huffman, and Ellen Magenheim. 2011. "When a Nudge Isn't Enough: Defaults and Saving Among Low-Income Tax Filers." *National Tax Journal*, 66.

Brown, Zack, Nick Johnstone, Ivan Haščič, Laura Vong, and Francis Barascud. 2012. "Testing the Effect of Defaults on the Thermostat Settings of OECD Employees." OECD 51.

Brunswik, Egon. 1952. *The conceptual framework of psychology. International encyclopaedia of unified science, 1:10*, Chicago:Univ. of Chicago Press.

Brunswik, Egon. 1956. *Perception and the representative design of psychological experiments.* . 2nd. ed., University of California Press.

Busemeyer, Jerome R., and In Jae Myung. 1988. "A new method for investigating prototype learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1): 3–11.

Camilleri, Adrian R., and Ben R. Newell. 2011. "When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms." *Psychonomic Bulletin & Review*, 18: 377–384.

Castillo, Geoffrey. 2020. "The Attraction Effect and Its Explanations." *Games and Economic Behavior*, 119: 123–147.

Cronqvist, Henrik, Richard H. Thaler, and Frank Yu. 2018. "When Nudges Are Forever: Inertia in the Swedish Premium Pension Plan." *AEA Papers and Proceedings*, 108: 153–158.

Crosetto, Paolo, and Alexia Gaudeul. 2016. "A Monetary Measure of the Strength and Robustness of the Attraction Effect." *Economics Letters*, 149: 38–43.

Debreu, Gerard. 1960. "Individual choice behavior: A theoretical analysis."

de Haan, Thomas, and Jona Linde. 2018. "'Good Nudge Lullaby': Choice Architecture and Default Bias Reinforcement." *Economic Journal*, 128(610): 1180–1206.

DeLosh, Edward L., Jerome R. Busemeyer, and Mark A. McDaniel. 1997. "Extrapolation: The sine qua non for abstraction in function learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(4): 968–986.

Dhami, Mandeep K., Ralph Hertwig, and Ulrich Hoffrage. 2004. "The Role of Representative Design in an Ecological Approach to Cognition." *Psychological Bulletin*, 130: 959–988.

Dietrich, Franz, and Christian List. 2016. "Reason-based choice and context-dependence: An explanatory Framework." *Economics and Philosophy*, 32(2): 175–229.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General*, 144(1): 114–126.

Dimant, Eugen, Gerben A. van Kleef, and Shaul Shalvi. 2020. "Requiem for a Nudge: Framing effects in nudging honesty." *Journal of Economic Behavior I& Organization*, 172: 247–266.

Dimara, Evanthia, Anastasia Bezerianos, and Pierre Dragicevic. 2017. "The Attraction Effect in Information Visualization." *IEEE Transactions on Visualization and Computer Graphics*, 23(1): 471–480.

Dinner, Isaac, Eric J. Johnson, Daniel G. Goldstein, and Kaiya Liu. 2011. "Partitioning default effects: Why people choose not to choose." *Journal of Experimental Psychology: Applied*, 17(4): 332–341.

Evangelidis, Ioannis, and Jonathan Levav. 2013. "Prominence versus Dominance: How Relationships between Alternatives Drive Decision Strategy and Choice." *Journal of Marketing Research*, 50(6): 753–766.

Evangelidis, Ioannis, Jonathan Levav, and Itamar Simonson. 2018. "The Asymmetric Impact of Context on Advantaged versus Disadvantaged Options." *Journal of Marketing Research*, 55(2): 239–253.

Farmer, George D., Paul A. Warren, Wael El-Deredy, and Andrew Howes. 2016. "The Effect of Expected Value on Attraction Effect Preference Reversals." *Journal of Behavioral Decision Making*, 30(4): 785–793.

Fodor, Jerry A. 1983. *The Modularity of Mind.* The MIT Press.

Forsgren, Mattias, Peter Juslin, and Ronald van den Berg. 2023. "Further perceptions of probability: In defence of associative models." *Psychological Review*, 130(5): 1383–1400.

Frederick, Shane, Leonard Lee, and Ernest Baskin. 2014. "The Limits of Attraction." *Journal of Marketing Research*, 51(4): 487–507. Publisher: SAGE Publications Inc.

Gerasimou, Georgios. 2016. "Partially dominant choice." *Economic Theory*, 61(1): 127–145.

Gigerenzer, Gerd. 2018. "The Bias Bias in Behavioral Economics." *Review of Behavioral Economics*, 5(3-4): 303–336.

Gomez, Yolanda, Víctor Martínez-Molés, Amparo Urbano, and Jose Vila. 2016. "The Attraction Effect in Mid-Involvement Categories: An Experimental Economics Approach." *Journal of Business Research*, 69(11): 5082–5088.

Hagmann, David, Emily H. Ho, and George Loewenstein. 2019. "Nudging out Support for a Carbon Tax." *Nature Climate Change*, 9.

Heath, Timothy B., and Subimal Chatterjee. 1995. "Asymmetric Decoy Effects on Lower-Quality Versus Higher-Quality Brands: Meta-Analytic and Experimental Evidence." *Journal of Consumer Research*, 22(3): 268.

Hedgcock, William M., Raghunath Singh Rao, and Haipeng (Allan) Chen. 2016. "Choosing to Choose: The Effects of Decoys and Prior Choice on Deferral." *Management Science*, 62(10): 2952–2976.

Hertwig, Ralph. 2017. "When to consider boosting: some rules for policy-makers." *Behavioural Public Policy*, 1(2): 143–161.

Hilgard, Sophie, Nir Rosenfeld, Jack Cao, Mahzarin Banaji, and David C. Parkes. 2019. "Learning Representations by Humans, for Humans."

Huber, Joel, John W. Payne, and Christopher P. Puto. 2014. "Let's be Honest about the Attraction Effect." *Journal of Marketing Research*, 51(4): 520–525. Publisher: SAGE Publications Inc.

Huber, Joel, John W Payne, and Christopher Puto. 1982. "Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis." *Journal of Consumer Research*, 9(1): 90.

Hursch, Carolyn J., Kenneth R. Hammond, and Jack L. Hursch. 1964. "Some methodological considerations in multiple-cue probability studies." *Psychological Review*, 71(1): 42–60.

Jachimowicz, Jon M., Shannon Duncan, Elke U. Weber, and Eric J. Johnson. 2019. "When and Why Defaults Influence Decisions: A Meta-Analysis of Default Effects." *Behavioural Public Policy*, 3(2): 159–186.

Johnson-Laird, P.N., Sangeet S. Khemlani, and Geoffrey P. Goodwin. 2015. "Logic, probability, and human reasoning." *Trends in Cognitive Sciences*, 19(4): 201–214.

Juslin, Peter, Sari Jones, Henrik Olsson, and Anders Winman. 2003*a*. "Cue abstraction and exemplar memory in categorization." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29: 924–941.

Juslin, Peter, Sari Jones, Henrik Olsson, and Anders Winman. 2003*b*. "Cue abstraction and exemplar memory in categorization." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29: 924–941.

Kaptein, Maurits C, Robin Van Emden, and Davide Iannuzzi. 2016. "Tracking the Decoy: Maximizing the Decoy Effect through Sequential Experimentation." *Palgrave Communications*, 2(1): 16082.

Karelaia, Natalia, and Robin M. Hogarth. 2008. "Determinants of linear judgment: A meta-analysis of lens model studies." *Psychological Bulletin*, 134(3): 404–426.

Krijnen, Job, David Tannenbaum, and Craig Fox. 2018. "Choice Architecture 2.0: Behavioral Policy as an Implicit Social Interaction." *Behavioral Science & Policy*, 3.

Król, Michał, and Magdalena Król. 2019. "Inferiority, Not Similarity of the Decoy to Target, Is What Drives the Transfer of Attention Underlying the Attraction Effect: Evidence from an Eye-Tracking Study with Real Choices." *Journal of Neuroscience, Psychology, and Economics*, 12(2): 88–104.

Lee, Chao-Feng, Shih-Chieh Chuang, Chou-Kang Chiu, and Kuo-Hao Lan. 2017. "The Influence of Task Difficulty on Context Effect - Compromise and Attraction Effects." *Current Psychology*, 36(3): 392–409.

Lichters, Marcel, Paul Bengart, Marko Sarstedt, and Bodo Vogt. 2017. "What Really Matters in Attraction Effect Research: When Choices Have Economic Consequences." *Marketing Letters*, 28(1): 127–138.

Logg, Jennifer. 2017. "Theory of Machine: When Do People Rely on Algorithms?" *SSRN Electronic Journal*.

Löfgren, Åsa, Peter Martinsson, Magnus Hennlock, and Thomas Sterner. 2012. "Are experienced people affected by a pre-set default option—Results from a field experiment." *Journal of Environmental Economics and Management*, 63(1): 66–72.

Maier, Maximilian, František Bartoš, T. D. Stanley, David R. Shanks, Adam J. L. Harris, and Eric-Jan Wagenmakers. 2022. "No evidence for nudging after adjusting for publication bias." *Proceedings of the National Academy of Sciences*, 119(31).

Mandl, Benjamin. 2022. "Cues, Beliefs, and Memory." PhD diss. Stockholm School of Economics, Stockholm.

Marewski, Julian N., and Lael J. Schooler. 2011. "Cognitive niches: An ecological model of strategy selection." *Psychological Review*, 118(3): 393–437.

Matysková, Ludmila, Brian Rogers, Jakub Steiner, and Keh-Kuan Sun. 2020. "Habits as Adaptations: An Experimental Study." *Games and Economic Behavior*.

Mazar, Nina, and Dilip Soman, ed. 2022. *Behavioral science in the wild. Behaviorally Informed Organizations*, Toronto, ON, Canada:University of Toronto Press.

McGuire, Joseph T., Matthew R. Nassar, Joshua I. Gold, and Joseph W. Kable. 2014. "Functionally Dissociable Influences on Learning Rate in a Dynamic Environment." *Neuron*, 84(4): 870–881.

McKenzie, Craig M. R. 2018. "Constructed Preferences, Rationality, and Choice Architecture." *Review of Behavioral Economics*, 5(3-4): 337–370.

Meredith, William. 1993. "Measurement invariance, factor analysis and factorial invariance." *Psychometrika*, 58: 525–543.

Mertens, Stephanie, Mario Herberz, Ulf J. J. Hahnel, and Tobias Brosch. 2021. "The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains." *Proceedings of the National Academy of Sciences*, 119(1).

Mertens, Stephanie, Mario Herberz, Ulf J. J. Hahnel, and Tobias Brosch. 2022. "Reply to Maier et al., Szaszi et al., and Bakdash and Marusich: The present and future of choice architecture research." *Proceedings of the National Academy of Sciences*, 119(31).

Morgenstern, Oskar, and John Von Neumann. 1953. *Theory of games and economic behavior.* Princeton university press.

Müller, Holger, Victor Schliwa, and Sebastian Lehmann. 2014. "Prize Decoys at Work — New Experimental Evidence for Asymmetric Dominance Effects in Choices on Prizes in Competitions." *International Journal of Research in Marketing*, 31(4): 457–460.

Nassar, Matthew R., Robert C. Wilson, Benjamin Heasly, and Joshua I. Gold. 2010. "An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment." *The Journal of Neuroscience*, 30(37): 12366–12378.

Natenzon, Paulo. 2019. "Random choice and learning." *Journal of Political Economy*, 127(1): 419–457.

Norton, Elyse H., Luigi Acerbi, Wei Ji Ma, and Michael S. Landy. 2019. "Human online adaptation to changes in prior probability." *PLOS Computational Biology*, 15(7): e1006681.

Padamwar, Pravesh Kumar, Jagrook Dawra, and Vinay Kumar Kalakbandi. 2019. "The Impact of Range Extension on the Attraction Effect." *Journal of Business Research*, S0148296319307830.

Palan, Stefan, and Christian Schitter. 2018. "Prolific.ac—A subject pool for online experiments." *Journal of Behavioral and Experimental Finance*, 17: 22–27.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research." *Journal of Experimental Social Psychology*, 70: 153–163.

Peterson, Cameron R., and Lee Roy Beach. 1967. "Man as an intuitive statistician." *Psychological Bulletin*, 68(1): 29–46.

Roe, Robert M., Jermone R. Busemeyer, and James T. Townsend. 2001. "Multialternative decision field theory: A dynamic connectionst model of decision making." *Psychological Review*, 108(2): 370–392.

Samuelson, William, and Richard Zeckhauser. 1988. "Status quo bias in decision making." *Journal of Risk and Uncertainty*, 1(1): 7–59.

Shafir, Eldar, Itamar Simonson, and Amos Tversky. 1993. "Reason-Based Choice." *Cognition*, 49(1-2): 11–36.

Shayna Rosenbaum, R., Alice S.N. Kim, and Stevenson Baker. 2017. "2.06 - Episodic and Semantic Memory." In *Learning and Memory: A Comprehensive Reference (Second Edition)*. . Second Edition ed., , ed. John H. Byrne, 87–118. Oxford:Academic Press.

Simonson, Itamar. 1989. "Choice Based on Reasons: The Case of Attraction and Compromise Effects." *Journal of Consumer Research*, 16(2): 158.

Smith, John P. 1969. "The effects of figurai shape on the perception of area." *Perception & Psychophysics*, 5(1): 49–52.

Soltani, Alireza, Benedetto De Martino, and Colin Camerer. 2012. "A Range-Normalization Model of Context-Dependent Choice: A New Model and Evidence." *PLOS Computational Biology*, 8(7): e1002607.

Sunstein, Cass R. 2017. "Nudges that fail." *Behavioural Public Policy*, 1(1): 4–25.

Szaszi, Barnabas, Anna Palinkas, Bence Palfi, Aba Szollosi, and Balazs Aczel. 2018. "A Systematic Scoping Review of the Choice Architecture Movement: Toward Understanding When and Why Nudges Work." *Journal of Behavioral Decision Making*, 31(3): 355–366. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.2035.

Szaszi, Barnabas, Anthony Higney, Aaron Charlton, Andrew Gelman, Ignazio Ziano, Balazs Aczel, Daniel G. Goldstein, David S. Yeager, and Elizabeth Tipton. 2022. "No reason to expect large and consistent effects of nudge interventions." *Proceedings of the National Academy of Sciences*, 119(31).

Tannenbaum, David, Craig R. Fox, and Todd Rogers. 2017. "On the Misplaced Politics of Behavioural Policy Interventions." *Nature Human Behaviour*, 1(7): 0130.

Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science*, 331(6022): 1279–1285.

Thaler, Richard H. 2015. "The power of nudges, for good and bad." *The New York Times*.

Thomas, de Haan, and Linde Jona. 2018. "'Good Nudge Lullaby': Choice Architecture and Default Bias Reinforcement." *The Economic Journal*, 128(610): 1180–1206.

Trueblood, Jennifer S., and Jonathan C. Pettibone. 2017. "The Phantom Decoy Effect in Perceptual Decision Making: Phantom Decoy in Perceptual Decision Making." *Journal of Behavioral Decision Making*, 30(2): 157–167.

Trueblood, Jennifer S., Scott D. Brown, and Andrew Heathcote. 2014. "The multiattribute linear ballistic accumulator model of context effects in multialternative choice." *Psychological Review*, 121(2): 179–205.

Trueblood, Jennifer S., Scott D. Brown, Andrew Heathcote, and Jerome R. Busemeyer. 2013. "Not Just for Consumers: Context Effects Are Fundamental to Decision Making." *Psychological Science*, 24(6): 901–908.

Tulving, Endel. 1972. "Episodic and semantic memory." In *Organization of memory.* , ed. Endel Tulving and Wayne Donaldson, 381–403. New York:Academic Press.

Yang, Sybil, and Michael Lynn. 2014. "More Evidence Challenging the Robustness and Usefulness of the Attraction Effect." *Journal of Marketing Research*, 51(4): 508–513. Publisher: SAGE Publications Inc.

Zhang, Jingjing, and Shawn P. Curley. 2018. "Exploring Explanation Effects on Consumers' Trust in Online Recommender Agents." *International Journal of Human–Computer Interaction*, 34(5): 421–432.