Topoi (2023) 42:799–807
https://doi.org/10.1007/s11245-023-09907-4

Algorithmic Nudging: The Need for an Interdisciplinary Oversight

Christian Schmauder[1] · Jurgis Karpus[1] · Maximilian Moll[2] · Bahador Bahrami[3,4,5,6] · Ophelia Deroy[1,4,7]

Accepted: 13 March 2023 / Published online: 3 April 2023

Abstract
Nudge is a popular public policy tool that harnesses well-known biases in human judgement to subtly guide people's decisions, often to improve their choices or to achieve some socially desirable outcome. Thanks to recent developments in artificial intelligence (AI) methods new possibilities emerge of how and when our decisions can be nudged. On the one hand, algorithmically personalized nudges have the potential to vastly improve human daily lives. On the other hand, blindly outsourcing the development and implementation of nudges to "black box" AI systems means that the ultimate reasons for why such nudges work, that is, the underlying human cognitive processes that they harness, will often be unknown. In this paper, we unpack this concern by considering a series of examples and case studies that demonstrate how AI systems can learn to harness biases in human judgment to reach a specified goal. Drawing on an analogy in a philosophical debate concerning the methodology of economics, we call for the need of an interdisciplinary oversight of AI systems that are tasked and deployed to nudge human behaviours.

Keywords Nudge · Algorithmic nudging · Bias in human judgement · Public policy · Explainable AI

1 Introduction

We are susceptible to external influences that make use of systematic biases in our judgement. Often we are not even aware of how predictable patterns in our thinking allow others to exert influence over what we choose and do. Nudge is a popular public policy tool that harnesses well-known biases in human judgement to subtly guide people's decisions. Usually this is done to achieve some socially desirable outcome (e.g., to increase the number of potential cadaveric organ donors in a society) or to help people attain outcomes that they would themselves agree to be best for them (e.g., adopt a healthy diet) but would not, left to their own devices, make an effort for.

As we begin to interact with artificial intelligence (AI) systems, do new possibilities emerge also of how and when our decisions can be nudged? On the one hand, nudging by AI can vastly improve daily human lives. Unlike nudges that target a society at large, which might work well for some people but not for others, AI systems can be deployed to develop and fine-tune personalized nudges, tailored to each individual separately. Healthcare is a particularly good example of a context in which, given the idiosyncrasy of patients, more effective personalized nudges could be developed thanks to big data and machine learning methods (Ruggeri et al. 2020). On the other hand, outsourcing the discovery and implementation of effective nudges to AI systems without proper oversight can have significant unintended negative side effects. This is especially so in the case of "black box" AI systems, the inner workings of which are

Christian Schmauder and Jurgis Karpus have contributed equally to this work.

✉ Jurgis Karpus
jurgis.karpus@lmu.de

1 Faculty of Philosophy, Philosophy of Science and Religious Studies, LMU Munich, Munich, Germany

2 Institute for Theoretical Computer Science, Mathematics and Operations Research, Universität der Bundeswehr München, Munich, Germany

3 Faculty of Psychology and Educational Sciences, LMU Munich, Munich, Germany

4 Munich Center for Neurosciences—Brain & Mind, Munich, Germany

5 Centre for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

6 Department of Psychology, Royal Holloway, University of London, London, UK

7 Institute of Philosophy, School of Advanced Study, University of London, London, UK

not easy to explain and monitor. While, in comparison to "human-designed" nudges, nudges developed by AI systems may be more effective in producing some desired outcome, the underlying human cognitive processes that they harness may not be fully understood.

In this paper, we unpack this concern by considering a series of examples. First, we provide examples of well-known cognitive biases in human judgement (Sect. 2) and discuss nudges that can harness these type of bias to steer people's decisions (Sect. 3). Then we discuss two recent case studies that demonstrate how AI systems can learn to harness biases in human judgment to achieve a specified objective (Sect. 4). Lastly, drawing on an analogy in a philosophical debate concerning the methodology of economics, we call for the need of an interdisciplinary oversight of AI systems that are tasked to nudge human behaviours (Sect. 5).

2 Biases in Human Judgement

It is by now amply evidenced that our thinking, for example, when we compare and evaluate various options presented to us, is often biased (Kahneman 2011). In analogy to optical illusions, biases in our thinking can be thought of as "cognitive illusions" (Pohl 2016). They are illusions of perception, judgement, or memory that distort our understanding of reality and sometimes make our thinking deviate from the normative principles of logic and also prudence. Importantly, the differences between reality and how we construe it are often systematic and, therefore, predictable. These "cognitive illusions" often occur involuntarily and are difficult to avoid. Below are some examples of well-known biases in human judgement that have been discovered, documented, and extensively studied to date (many more are reviewed by Kahneman 2011 and by contributing authors in Pohl 2016; at the time of writing this article, the Wikipedia page on cognitive biases listed 251 of them[1]).

2.1 The Conjunction Fallacy

Sometimes we judge the possible occurrence of a conjunction of two events to be more probable than the possible occurrence of one of those events, irrespective of the occurrence of the other. This contradicts logic. But a possible occurrence of a conjunction of two events can, in some circumstances, be more familiar to us compared to the possibility of occurrence of only one of those events. The canonical example was introduced by Tversky and Kahneman (1983). Consider Linda, a graduate in philosophy who is known to be deeply concerned with social justice and has recently participated in anti-nuclear demonstrations. Is it more likely that Linda is a banker and is active in the feminist movement (proposition *a*) or that Linda is a banker (proposition *b*)? When people were asked to rank these propositions among various others (e.g., that Linda is a teacher in an elementary school) in terms of their overall likelihood, many ranked the conjunctive proposition *a* above *b*, despite the fact that *b* must be at least as likely as *a* (the set of bankers who are active in the feminist movement is a subset of the set of bankers). While the example itself is somewhat dated (today, we rarely refer to feminism as the feminist "movement" and, hopefully, many more of us have become feminists since the 1980s), it shows how quick and intuitive associations that we often rely on can lead us to commit a logical fallacy.

2.2 The Illusion of Control

People prefer raffles in which they can choose a number of their liking for a random draw to those in which this number is given to them, despite the fact that the objective chances of winning are the same in both scenarios (Langer 1975). Relatedly, when asked about driving safety, many people rated the probability of being involved in a car accident to be lower when they are in the driver's seat compared to when they are a passenger (McKenna 1993). These examples show that we tend to (falsely) think that by simply taking an action we can exert control over the outcomes of events that are purely probabilistic.

2.3 Anchoring

Our evaluations of options can be predictably swayed by reference points—starting points for our thinking—even when there are no rational grounds to base our judgements on such "anchors." Tversky and Kahneman (1974) asked people to estimate the percentage of African countries among those in the United Nations. Before producing their estimate, one group of participants was exposed to 10% as the starting point for their thinking. Another group was exposed to 65%. The median reported estimates in the two groups were 25% and 45% respectively (the true value is 28%). More strikingly, Strack and Mussweiler (1997) asked participants in one group of people in their experiment whether they thought Mahatma Gandhi died before or after the age of 9. They asked another group whether they though he died before or after the age of 140. Following this, participants in both groups estimated Gandhi's actual age at the time of death. Even though the two anchors were clearly nonsensical, the mean reported estimates were 50 and 67 in the two groups respectively (the correct answer is 78).

[1] https://en.wikipedia.org/wiki/List_of_cognitive_biases.

2.4 The Hindsight Bias

We tend to think that the outcomes of probabilistic events that we actually experienced were more predictable than they actually were before those events took place. In a famous study of this "I-knew-it-all-along" effect, Fischhoff and Beyth asked people to judge the likelihood of various possible (at the time, future) outcomes of US President Richard Nixon's planned state visits to China and the USSR in the early 1970s. The same people were later asked to remember or reconstruct their earlier predictions after the events took place. People's remembered probabilities that they predicted were higher for those outcomes that participants believed to have actually occurred. In other words, people were less surprised about the actual outcomes of President Nixon's visits after those events had happened than they were before those events took place (Fischhoff and Beyth 1975).

2.5 The Outcome Bias

Somewhat relatedly, when we evaluate the goodness of our past decisions concerning (at the time, future) probabilistic events, we tend to overweight the importance of the actual outcomes of those probabilistic events after any uncertainty concerning them has been resolved (Baron and Hershey 1988). For example, many of us have the tendency to judge the goodness of a decision to purchase a lottery ticket based on the subsequent outcome of that lottery. However, the outcome of the lottery is, of course, unknown at the time of making the decision. In our after-the-fact judgement, we tend to underweight the importance of the true odds of winning and the potential consequences of losing that were known to us at the time of purchase.

3 Nudge

In their seminal book, Thaler and Sunstein described nudge as 'any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives' (2008, p. 6). Earlier they had given the initial impetus to the idea of nudge arguing that, since systematic biases in human judgement make people's choices imperfect, many day-to-day decisions could be improved with the help of subtle, non-coercive interventions (Sunstein and Thaler 2003). Based on their definition, a policy intervention is a nudge if it fulfils three criteria. First, a nudge concerns the design of choice architecture—broadly put, any feature of the context in which decision-makers make choices. Second, a nudge does not take away any of the available choice options from decision-makers. Coercive interventions, such as outright bans, are not nudges. Third, a nudge does not significantly alter the economic incentives to decision-makers. Taxes on sugary drinks or subsidised public transport, while being able to alter people's behaviour in predictable ways, are not nudges.

Several refinements of the original definition have been proposed later on. For example, Hansen (2016, p. 158) stressed the connection to the irrationality of biases in human judgement: 'a nudge is a function of any attempt at influencing people's judgement, choice or behaviour in a predictable way that is made possible because of cognitive boundaries, biases, routines and habits in individual and social decision-making posing barriers for people to perform rationally in their own declared self-interests and which works by making use of those boundaries, biases, routines, and habits as integral parts of such attempts.' Since this description does not mention economic incentives, interventions that mix economic incentives with more subtle psychological techniques, for example, a small tax on the use of plastic bags at a supermarket, can presumably still count as nudges. However, since Hansen's definition stresses the *rationality* failures in human judgement, it can be thought of as a narrower definition of nudge than Thaler and Sunstein's (Berthet and Ouvrard 2019). An educational campaign that informs people about the risks of contracting sexually transmitted diseases in order to promote the use of contraceptives could be classed as nudge according to Thaler and Sunstein, but not Hansen.

3.1 Beneficiary

Nudges are *pro-self*, *pro-social*, or *pro-nudger*, depending on who the ultimate beneficiary of a nudge is. If the sole purpose of a nudge is to benefit the nudged person, it is *pro-self* (Congiu and Moscati 2022). Such nudges appear to be the least controversial and the most accepted by society at large (Hagman et al. 2015). They are typically used in situations in which, due to some cognitive bias in human judgement, people do not choose the option that is in their own best interest. For example, a nudge that results in an increase in people's retirement savings can be seen as promoting the nudged people's own future well-being.

Pro-social nudges do not or only partially benefit the nudged person. Their primary goal is to increase the welfare of society at large. Nudges that promote tax compliance, energy conservation, or charitable donations may or may not be *pro-self*, but they are *pro-social*. For example, households can be effectively nudged into consuming less energy by regularly providing them information on how their energy consumption compares to that of their close neighbours (Allcott and Rogers 2014), or by setting the default to a green energy provider, when people move into their new homes, rather than a possibly cheaper, "greyer" alternative (Sunstein 2016).

While effective in producing behavioural change, this form of nudging can come at the cost of unintended, problematic side effects. In the case described above, it was subsequently found that people who tend to stick with the default green energy provider are overall poorer compared to those who tend to switch their energy provider to a cheaper, "greyer" alternative (Sunstein 2016). That raises a moral dilemma for social planners: is it okay to use subtle, non-coercive tools to steer people's decisions to attain a state of affairs in which the poorer half of our society pays more for tackling climate change compared to the more wealthy half?

Some marketing techniques work just like nudges. Online retailers target our tendency to social conformity when they advertise products as "frequently bought," employ decoys when displaying inferior products next to the product that a consumer is interested in buying, or use anchoring when they display a list price of the product as a reference, while the product is actually sold at a discount (Congiu and Moscati 2020). Sometimes these nudging methods can benefit both the nudger and the person being nudged. This gives room for nudges that are part *pro-nudger*, part *pro-self*, or part *pro-nudger*, part *pro-social*, where private companies, as well as their customers or society at large benefit from the same nudge (Congiu and Moscati 2022). One of the case studies of AI-powered nudging that we will discuss later will concern this hybrid type of nudge.

3.2 Target Thought Processes

Nudges also vary based on the cognitive process that they are designed to influence. The dual-process theory in psychology suggests that our brains work at two different levels (Kahneman 2003; Julmi 2019). System 1 refers to implicit cognitive processes closely related to intuition. These processes are not those that we customarily call "thinking." The "thought" operations here happen automatically and often subconsciously. They are fast and difficult to control. In contrast, System 2 describes reflective processes that are performed consciously in a controlled and planned manner. Compared to the workings of System 1, System 2 is slow, effortful, and people are consciously aware of it when it is engaged. Not everyone agrees that this theorized dichotomy is right, but the distinction is useful to consider two contrastive ways of thinking: one that is fast, e.g., when it relies on intuitions, and the other—more reflective (see, for example, Mercier and Sperber 2011).

In situations where intuition dominates, our brains need to act fast, which implies that they need to rely on heuristics—quick rules of thumb—to evaluate a situation at hand and to make a quick decision. These rules of thumb allow brains to act quickly, but they come at the cost of biasing our "thinking" and our choosing. Nudges that target intuitive processes steer people's decisions without engaging the reflective processes of their cognition. Nudges that engage the reflective processes target slower, more deliberate domain of thinking. Examples of the latter are educational nudges that provide decision-makers with relevant information—information that they would not otherwise consider—in order to allow them to make better informed decisions, slowly, consciously, and deliberatively. In the rest of this paper, we will be concerned with the deployment of AI systems to develop and implement nudges that target the less reflective domain of thinking—those that are harder to detect by people who are nudged.

4 Nudge in Human-AI Interactions

As more aspects of our daily lives are digitized, more data about our preferences, choices, and beliefs can be amassed and studied. According to some estimates, two average days in 2010 saw as much data created as during the entire history of humanity up to 2003 (Siegler 2010). In a single minute in 2019, people collectively watched the equivalent of 700,000 hours of video on Netflix, sent 188 million emails, and entered 3.8 million search queries on Google (Desjardins 2019). Digital technologies are transforming how we interact with others, work, and consume and process information (Nadkarni and Prügl 2021). Our interaction with computers and AI systems also creates new possibilities for how and when our behaviours can be nudged (Caraban et al. 2019). Below we discuss two case studies that demonstrate how AI systems can learn to harness biases in human judgment to reach their goal.

4.1 The Bandit Task Experiment

Dezfouli et al. (2020) conducted three experiments, in each of which they trained an AI agent to develop strategies to subtly sway human decision-makers' choices. In one of their experiments (the Bandit Task) human participants repeatedly chose between two lotteries that would either yield or not yield a reward to them in any trial (iteration) of the game. Throughout 100 trials, each lottery yielded a reward exactly 25 times, but participants did not know on which trials which lottery (if any, or perhaps both) would yield them a reward. If for each participant the rewards for both lotteries were distributed randomly across the 100 trials, we should expect people, on average, to choose either lottery 50% of the time.

With this setup, Dezfouli and colleagues developed and trained an AI agent to distribute rewards for each lottery across the 100 trials so as to sway human decision-makers into favouring one lottery over the other. At the end of its learning phase, the AI agent could successfully nudge human decision-makers to choose the specified "target" lottery 70%

of the time—a statistically significant shift from the 50% baseline. Moreover, the agent learned to dynamically adapt its optimal strategy to each individual human participant in the experiment by observing that participant's decision-making style during the early trials of the task. Strategies that were optimal to sway participants who engaged in an extensive trial-and-error exploration of the two lotteries in early trials of the game were different from those that were optimal to sway participants who, at the start of the game, stubbornly stuck with the first lottery that yielded a reward. This shows that AI systems can indeed develop and fine-tune personalized nudges tailored to each individual decision-maker separately.

One well-documented bias in human judgement that the developed AI agent could exploit is the primacy effect. When we meet new people, try new experiences, or evaluate new products, the first impressions we form about them tend to stick. This is well known to advertisers. Extensive marketing campaigns are often launched before new products are released into the market to pre-emptively create good first impressions of them (Murphy et al. 2006). Evidence for the primacy effect comes in various "flavours" and has been reported in different contexts. For example, the first wine one samples during a wine-tasting session tends to be chosen significantly more often than other wines at the end of the session (Mantonakis et al. 2009). Shteingart et al. (2013) describe an experiment in which participants repeatedly chose between safe (no risk, mediocre gain) and risky (high risk, high potential gain) options to win money. Among participants who chose the risky option in their first trial, those who won subsequently chose the risky option significantly more often than those who lost (on average, 47% and 31% of the time, respectively).

In the report of the results of their study, Dezfouli and colleagues illustrate how their developed AI agent was able to identify and exploit the primacy effect in some of the recruited human participants' choices observed in early trials of the game. However, the AI agent did not "assume" the primacy effect to occur across its interactions with *all* human participants in the experiment and, hence, was able to use different strategies for different participants to optimize its overall ability to sway as many people's decisions as possible across the board.

4.2 The Advising Game Experiment

AI-powered recommendation engines that provide personalized advice on which movies to watch, which books to read, or which websites to visit, are a good example of automated advice-giving systems that we already use today. Many of these systems are built on the assumption that the better their recommendations are, the more we will rely on them to make personal decisions (Schrage 2020). Conversely, our use of these systems can be a reliable indicator that they fulfil their promise of issuing good recommendations to us.

In theory, this gives a neat and practically convenient result. An automated advice-giving system tasked to attract and retain its human users will simultaneously win business for the service provider and give its users what they truly want—good recommendations to make better-informed decisions. In this light, our continued and increasing use of recommendation engines serves as evidence that we watch more movies that we truly like, read more books that we find truly interesting, and spend less time browsing the web to identify websites that contain information that we are truly after.

This reasoning makes good sense in a one-to-one relationship between an adviser and its client, where the client simply chooses whether to use the adviser's issued recommendations (to inform the client's decisions) or not. However, the assumptions that underlie this reasoning break down when multiple advisers compete for a single client's attention. This is so because, in a competitive market, advisers that are tasked to attract and retain their clients will care not only about the quality of their issued advice, but also about winning clients away from the competition.

To illustrate this point, consider a simple game in which we place bets on sides in a series of football matches and we can turn to an adviser to inform our placement of those bets. Suppose that our ultimate and only goal is to maximize our winnings (for example, we don't care about the mere buzz associated with placing bets, which can at times be pleasurable in and of itself). Suppose also that we have no information whatsoever about the football teams involved and, therefore, equate our prospect of picking the winning team in any match with a 50% chance. In this setting, if we have any reason to believe that the adviser that we can turn to is better informed about the odds of winning than mere chance, it makes perfect sense for us to follow that adviser. If we can thereby make a profit, it would even be wise for us to subscribe to this adviser's service for a fee.

Suppose now that a second adviser enters the market. This adviser has the exact same information about the odds as the adviser we already follow. But we do not know this. If the second adviser communicates the exact same recommendations as our original adviser (which they should if they wish to communicate their advice truthfully), we will have no reason to switch to them. This holds irrespective of how any single bet that we place actually plays out in any given match because the advisers will tell us the same thing. The only way the new adviser will be able to attract our attention is by saying something different. Knowing that our chosen oracle (our initially followed adviser) will support the team favoured by the current odds, the newcomer can fervently support the underdog and hope that an unexpected outcome of a match, if it occurs, will upset our trust in our present

oracle and will lead to us choosing them. At some point a low probability event (an underdog winning) is bound to happen, since the outcome of any football match is, after all, a probabilistic event.

Several recent studies confirmed this prediction empirically. For example, Kurvers and colleagues (2021) conducted a series of experiments using a game similar to the one above. They found that dishonest advisers—advisers that did not always communicate the odds of winning truthfully—consistently outperformed honest advisers in competition for attracting a human client. When people played the role of the advisers in the game, they quickly learned to use such tactics too (Hertz et al. 2018), as did AI-powered algorithmic advisers in tests with (simulated) human clients (Moll et al. Forthcoming).

One reason for why dishonest advice-giving strategies successfully attract human clients is the outcome bias in human judgement that we discussed earlier (Kurvers et al. 2021). Because of our tendency to overweight the importance of the actual outcomes of probabilistic events, even when we know that the outcome of an event is probabilistic in nature, whoever happens to correctly guess that outcome immediately attracts our attention. AI-powered advisers can uncover and exploit this tendency to achieve their goals. Importantly, this example shows that, without game-theoretic methods to test this hypothesis empirically and without the knowledge of cognitive science and experimental psychology research to interpret behavioural findings from such studies, dishonest strategies in algorithmic advice-giving can emerge without anyone's malicious intent and they may never be noticed by unsuspecting developers, providers, and end users of automated systems.

5 The Need to Look Under the Hood

A nudge always serves a specific purpose. In the role of a choice architect, a social planner may be interested in developing nudges that would help people adopt healthier, environmentally sustainable diets, take up physical exercise, conserve energy, or save for retirement. Given some such benevolent goals, there may, however, be many ways to achieve them. Knowing exactly how a nudge does that is helpful and important. This is especially so if a nudge, while successful in bringing about the desired outcome, is found to produce an unintended, unwelcome side effect. Once some such side effect has been brought into light, without a good understanding of what exactly causes it and how, the only remedy to the problem might be to abandon the implemented nudge altogether, or implement a new, more complicated nudge on top of the first one to cancel the undesired effects of the former. If, on the other hand, one understands the mechanism through which the nudge produces the unwanted effect, one can work on improving or fixing that nudge instead. The need to be able to "look under the hood" of a mechanism employed to attain the desired objective is not limited to the outsourcing of the development and implementation of nudges to "black box" AI systems.

5.1 A Lesson from Economics

In a famous philosophical debate concerning the methodology of economics in the middle of last century Friedman (1953) argued that, while normative economics is concerned with values and what ought to be, for example, what constitutes a just and fair distribution of wealth in a society, economics as a positive science deals merely with empirical facts and testable predictions. As social scientists, economists, according to Friedman, are primarily in the business of developing theories that make accurate and useful predictions. The predictions of these theories can then inform policies developed by social planners, who, among other things, are concerned with normative questions about how social affairs ought to be conducted and regulated. According to this instrumentalist view of economics, a theory is only good, from the perspective of a social scientist, insofar as it produces accurate predictions. Any other aspect of the theory is essentially irrelevant. As long as the theory's predictive power is uncompromised, the "inner workings" of the theory itself, for example, the realism of axioms and assumptions on which the theory is built, do no matter.

Friedman's view was and still is influential, but has also been criticized. Hausman (1994) challenged Friedman's thesis by constructing an analogy as a counterargument. He considered what makes a good car. Ultimately, a good car is one that performs well in fulfilling its purpose—it is a safe and reliable mode of transport. Extending Friedman's idea to the evaluation of cars, we could determine whether some car is good simply by observing how well it performs in a test drive. Any other aspect of the car is presumably irrelevant for assessing its quality. This view, Hausman argues, is short-sighted. Certainly nobody who plans to purchase a used car would accept this suggestion as true. Test drives are undoubtedly important, but they are only a part of a thorough evaluation of a car. A potential buyer of a used car will want to know not only how well it performs in a test drive on some given day, but also how well it would perform in different circumstances (for example, when it rains or when it is freezing cold). A prudent buyer will demand to see a report of the most recent thorough inspection of the car's components. They may even hire a trustworthy mechanic to take another look under the car's hood to assess their present condition.

Hausman argues that, similarly to the case in his analogy, when we evaluate a theory, a thorough inspection of its "inner workings" (that is, in addition to the theory's

performance in making decent predictions) is not merely helpful, but often also necessary. This becomes evident when, applied to novel circumstances, the theory fails to produce an accurate prediction. Upon encountering such an event, it would be hasty to ditch the theory outright deeming it utterly useless. One could instead attempt to fix it by inspecting which of the theory's many assumptions are most likely at fault. If a closer scrutiny reveals that some of the theory's assumptions are clearly false in the circumstances under which it failed to produce a sufficiently accurate prediction, those particular assumptions might need to be relaxed or corrected. Returning to Hausman's analogy, it would be crazy if, whenever our cars broke down, we would immediately replace them with new ones. It makes much more sense to tow a broken down car to a service station and ask a specialist to take a look under its hood in order to determine whether a fix or a replacement of one of its components might do.

We can extend this debate from the philosophy of economics to the context in which we evaluate the performance of a "black box" system that is tasked to produce a desired outcome. According to Friedman, there is no need to question the inner workings of such a system, so long as it succeeds in producing that desired outcome. According to Hausman, because the system *can* unexpectedly break down, for example, when it is applied to novel circumstances or when it produces a hitherto unpredicted side effect, disregarding the actual processes by which it generates the desired outcome is not a good idea. This way we can extend Hausman's argument to the use of "black box" AI systems that may be employed to develop and implement nudges to steer people's decisions. In addition to monitoring how well such systems achieve the goals that we set for them, we want to be able to inspect and understand their inner workings in order to know what to do when things do not go as planned. Put simply, we need to be able to look under their hood.

5.2 A Call for an Interdisciplinary Oversight

A serious drawback of blindly outsourcing the development and implementation of nudges to "black box" AI systems is that the ultimate reasons for why such nudges work, that is, the underlying human cognitive processes that they harness, will often be unknown. A personalized nudge developed by an AI system might be admirably effective in producing some desired outcome. But which biases in human judgement (if any are at play) such a nudge harnesses will be hidden in the system's inner "black box" workings. After all, an AI system may not "knowingly" utilize any bias in human judgement—it would merely do what works to attain its specified objective. That will be a problem when it comes to foreseeing and subsequently dealing with unintended side effects that the implemented nudge might eventually produce.

Consider again the case of algorithmic advising that we discussed in Sect. 4.2. Suppose a developer of an automated advising system works under the assumption that people's use of the system will correlate with the quality of the system's communicated advice to them (an assumption that actually is true in a one-to-one relationship between an adviser and its client). In that case, it will be sufficient—as well as practically convenient and mutually beneficial to all parties involved—to "instruct" the advice-issuing system simply to attract and retain its human users. If, at a later stage, it is uncovered that, contrary to the prediction, a highly popular automated system communicates advice to its human users untruthfully, we might have no idea why that is the case. Crucially, without a good theory—in this case, without the knowledge of the existence and a thorough understanding of the outcome bias in human judgement—we will lack ideas for how to improve such systems, and how to regulate their deployment and use.

Recall our earlier discussion of the default rule nudge that effectively steered people's decision to choose a green energy provider for their new homes (Sect. 3.1). The unintended side effect of this, at first sight, successful nudge was that people who tended to stick with the suggested default (that is, on whom the nudge was most effective) were overall poorer compared to people who switched their energy provider to a cheaper, "greyer" alternative. In order to identify the most appropriate response to the discovery of this unwelcome outcome, it is important to understand the underlying reasons for why the nudge worked in this context to begin with. Default rule nudges can be effective for a number of different reasons (Beraldo and Karpus 2021). If people stick with the default option because they deem this option to be implicitly recommended to them by policy-makers, the best response might be to investigate why the richer part of society tends to ignore this implicit recommendation. If, on the other hand, the default stuck for those people who continually delayed making a decision about their energy provider because they had other, more important decisions to focus on in their daily lives, the most appropriate response might be to abandon the introduced nudge altogether and to create opportunities and time for people to consider the importance of their choice and our communal need to tackle global warming.

The crux of the problem lies not merely in the need to make "black box" AI systems explainable, in the sense that they should indicate which *known* human cognitive processes they learn to harness to achieve their specified goals, but in the need to monitor which potentially *not-yet-known* human cognitive processes they can learn to harness. Cognitive scientists and psychologists investigate, explain, and try to understand human cognition and behaviour. In this

light, they will be valuable and, indeed, necessary colleagues to computer scientists and software engineers in developing and, more importantly, monitoring AI systems that are tasked and deployed to nudge human decision-makers "in the field."

6 Conclusion

Algorithmically personalized nudging has the potential to vastly improve daily human lives. As we argued, however, blindly outsourcing the development and implementation of nudges to "black box" AI systems comes with a serious drawback. When a human expert develops a nudge, they are informed by pre-existing theory and knowledge of human cognitive processes and biases in human judgement. At the very least, this theory informs our understanding of why developed nudges work and how to deal with any of their unintended side effects. When an AI system develops a nudge, it may not be informed by any pre-existing theory, making it harder to understand why its developed nudge actually works. Furthermore, if the AI system learns to harness some yet unknown, undocumented bias in human judgement, such a theory may not exist at all, even in principle. That makes it particularly difficult to foresee potential pitfalls and to subsequently fix unwelcome consequences of nudging if and when they come to light.

Large technology companies are increasingly under pressure to regulate themselves, for example, by developing and imposing codes of ethical conduct for their employees to follow (Nemitz 2018; Denning 2020). The obligation to use and leverage the expertise of cognitive scientists and psychologists in cases where developed "black box" AI systems can profoundly impact decisions made by their human users should be added to these efforts. From a governmental perspective, requiring the developers of AI systems to consult experts on human judgement and decision-making would increase the accountability of companies for the effects of their marketed AI-powered tools. From a societal perspective, knowing that the "black box" systems are routinely reviewed and monitored by experts on human cognition could boost people's trust in them (which resonates with the guidelines for trustworthy AI recently presented to the European Commission by the High-Level Expert Group on Artificial Intelligence 2019). The same would be true of their acceptance and use.

Author Contributions Christian Schmauder and Jurgis Karpus contributed equally to the development of this paper.

Funding Open Access funding enabled and organized by Projekt DEAL. Bahador Bahrami was supported by the Humboldt Foundation and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (819,040—acronym: rid-O). Ophelia Deroy was supported by the NOMIS Foundation and the Research Council of Norway project "Warring with Machines" at the Peace Research Institute Oslo (PRIO).

Declarations

Conflict of interest The authors have no competing interests to disclose.

References

Allcott H, Rogers T (2014) The short-run and long-run effects of behavioral interventions: experimental evidence from energy conservation. Am Econ Rev 104:3003–3037. https://doi.org/10.1257/aer.104.10.3003

Baron J, Hershey JC (1988) Outcome bias in decision evaluation. J Pers Soc Psychol 54:569–579

Beraldo S, Karpus J (2021) Nudging to donate organs: do what you like or like what we do? Med Health Care Philos 24:329–340. https://doi.org/10.1007/s11019-021-10007-6

Berthet V, Ouvrard B (2019) Nudge: towards a consensus view? Psychol Cogn Sci Open J 5:1–5. https://doi.org/10.17140/PCSOJ-5-143

Caraban A, Karapanos E, Gonçalves D, Campos P (2019) 23 ways to nudge: a review of technology-mediated nudging in human-computer interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 503. https://doi.org/10.1145/3290605.3300733

Congiu L, Moscati I (2020) Message and environment: a framework for nudges and choice architecture. Behav Public Policy 4:71–87. https://doi.org/10.1017/bpp.2018.29

Congiu L, Moscati I (2022) A review of nudges: definitions, justifications, effectiveness. J Econ Surv 36:188–213. https://doi.org/10.1111/joes.12453

Desjardins J (2019) What happens in an internet minute in 2019? Visual capitalist. https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/

Denning S (2020) Why big tech should regulate itself. Forbes. https://www.forbes.com/sites/stevedenning/2020/08/02/why-big-tech-should-regulate-itself/

Dezfouli A, Nock R, Dayan P (2020) Adversarial vulnerabilities of human decision-making. PNAS 117:29221–29228. https://doi.org/10.1073/pnas.2016921117

Fischhoff B, Beyth R (1975) I knew it would happen: remembered probabilities of once—future things. Organ Behav Hum Perform 13:1–16. https://doi.org/10.1016/0030-5073(75)90002-1

Friedman M (1953) The methodology of positive economics. Essays in positive economics. The University of Chicago Press, Chicago

Hagman W, Andersson D, Västfjäll D, Tinghög G (2015) Public views on policies involving nudges. Rev Philos Psychol 6:439–453. https://doi.org/10.1007/s13164-015-0263-2

Hansen PG (2016) The definition of nudge and libertarian paternalism: does the hand fit the glove? Eur J Risk Regul 7:155–174. https://doi.org/10.1017/S1867299X00005468

Hausman DM (1994) Why look under the hood? The philosophy of Economics: an anthology. Cambridge University Press, Cambridge

Hertz U, Palminteri S, Brunetti S, Olesen C, Frith CD, Bahrami B (2018) Neural computations underpinning the strategic management of influence in advice giving. Nat Commun 8:2191. https://doi.org/10.1038/s41467-017-02314-5

High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels

Julmi C (2019) When rational decision-making becomes irrational: a critical assessment and re-conceptualization of intuition effectiveness. Bus Res 12:291–314. https://doi.org/10.1007/s40685-019-0096-4

Kahneman D (2003) Maps of bounded rationality: psychology for behavioral economics. Am Econ Rev 93:1449–1475. https://doi.org/10.1257/000282803322655392

Kahneman D (2011) Thinking, fast and slow. Allen Lane, Bristol

Kurvers RHJM, Hertz U, Karpus J, Balode MP, Jayles B, Binmore K, Bahrami B (2021) Strategic disinformation outperforms honesty in competition for social influence. iScience 24:103505. https://doi.org/10.1016/j.isci.2021.103505

Langer EJ (1975) The illusion of control. J Pers Soc Psychol 32:311–328

Mantonakis A, Rodero P, Lesschaeve I, Hastie R (2009) Order in choice: effects of serial position on preferences. Psychol Sci 20:1309–1312. https://doi.org/10.1111/j.1467-9280.2009.0245

McKenna FP (1993) It won't happen to me: unrealistic optimism or illusion of control? Br J Psychol 84:39–50. https://doi.org/10.1111/j.2044-8295.1993.tb02461.x

Mercier H, Sperber D (2011) Why do humans reason? Arguments for an argumentative theory. Behav Brain Sci 34:57–74. https://doi.org/10.1017/S0140525X10000968

Moll M, Karpus J, Bahrami B (Forthcoming) (eds) Do artificial agents reproduce human strategies in the advisers' game? Operations Research Proceedings. Springer

Murphy J, Hofacker C, Mizerski R (2006) Primacy and recency effects on clicking behavior. J Comput-Mediat Comm 11:522–535. https://doi.org/10.1111/j.1083-6101.2006.00025.x

Nadkarni S, Prügl R (2021) Digital transformation: a review, synthesis and opportunities for future research. Manag Rev Q 71:233–341. https://doi.org/10.1007/s11301-020-00185-7

Nemitz P (2018) Constitutional democracy and technology in the age of artificial intelligence. Philos Trans Royal Soc A 376:20180089. https://doi.org/10.1098/rsta.2018.0089

Pohl RF (2016) Cognitive illusions. In: Pohl RF (ed) Cognitive illusions: intriguing phenomena in judgement, thinking and memory. Psychology Press, Hove

Ruggeri K, Benzerga A, Verra S, Folke T (2020) A behavioral approach to personalizing public health. Behav Public Policy. https://doi.org/10.1017/bpp.2020.31

Schrage M (2020) Recommendation engines. The MIT Press, Cambridge

Shteingart H, Neiman T, Loewenstein Y (2013) The role of first impression in operant learning. J Exp Psychol 142:476–488. https://doi.org/10.1037/a0029550

Siegler MG (2010) Eric Schmidt: every 2 days we create as much information as we did up to 2003. TechCrunch, San Francisco

Strack F, Mussweiler T (1997) Explaining the enigmatic anchoring effect: mechanisms of selective accessibility. J Pers Soc Psychol 73:437–446. https://doi.org/10.1037/0022-3514.73.3.437

Sunstein CR, Thaler RH (2003) Libertarian paternalism is not an oxymoron. U Chi L Rev 70:1159–1202. https://doi.org/10.2307/1600573

Sunstein CR (2016) The Ethics of Influence: government in the age of behavioral science. Cambridge University Press, Cambridge

Thaler RH, Sunstein CR (2008) Nudge: improving decisions about Health, Wealth, and happiness. Yale University Press, New York

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185:1124–1131

Tversky A, Kahneman D (1983) Extensional versus intuitive reasoning: the conjunction fallacy and probability judgment. Psychol Rev 90:293–315. https://doi.org/10.1037/0033-295X.90.4.293