# Robot nudgers. What About Transparency?

**Stefano Calboli** (Centre for Ethics, Politics and Society)[1]

## Abstract

Robot nudgers – i.e. robots who employ "nudges" to steer users toward targeted behaviours – are a concrete reality nowadays (Ali Mehenni *et al.* 2020; Hang *et al.* 2021). Albeit robot nudgers look like a promising technology for making individuals and society better off (Borenstein & Arkin 2015), some ethically relevant questions in programming them have been so far under-examined.

The paper aims to contribute to filling this gap, identifying two ethical issues concerning nudges' transparency relevant when robots step into the shoes of nudgers. I proceed as follows.

The paper begins by outlining what policy tools can be considered nudges (§1) and why scholars advocate for making their implementation transparent in order to shield persons' decision-making autonomy (see Ivanković & Engelen 2019; Wilkinson 2012) (§2).

Therefore, I focus on the still unripe literature on robot-nudging (Rodogno 2020; Borenstein & Arkin 2016) and, in light of it, I properly frame ethical issues concerning nudges' transparency in human-robot interactions (§3). In Section 4, I discuss two ethically relevant points concerning transparency in robot-nudging so far overlooked. First, Robot nudgers - in contrast with human-nudgers - are potentially able to customize the kind of transparency granted to a specific user. Second, robot nudgers are able to monitor the impact of any feasible mixes of nudges and transparency on the effectiveness of the nudges in steering decision-makers. In both cases, ethically relevant questions emerge.

I conclude by advocating for the involvement of ethicists in robot nudgers' programming at an early stage, in line with an integrative approach to social robotics (§5).

## I. Introduction

At least since the publication of the book '*Nudge: Improving Decisions About Health, Wealth*' by Thaler and Sunstein in 2008 (last edition in 2021), nudges should, and in fact often are (OECD 2017), valuable parts of policymakers' toolbox.

To understand what nudges are and why they are considered no-conventional policy tools, we should briefly delve into the theoretical background that has made possible their conception.

The research carried out by Herbert Simon (1955) on human bounded rationality has provided fertile ground for the coming of what has been called the "behavioural revolution" and the development of approaches aimed at modeling decision-making, taking into account a model truest to *human cognition* as it is.

[1] This work has been supported by Fundação para a Ciência e a Tecnologia, prot. UI/BD/152568/2022.

Such approaches contrast with the one adopted in neoclassical economics in which *homo oeconomicus* (HO henceforth) is the model. HO is based on deliberately (see Levine 2020) highly idealized assumptions. Relevant to the purpose of the present paper, HO is featured by the following three traits.

First, HO is *perfectly rational and has infallible cognitive abilities*. For instance, he or she is able to evaluate the expected utility of lotteries with no chance of failure.

Second, HO has *perfect willpower and practical abilities*. Let us say that the agent decides to participate in a lottery; if so, afterthoughts, procrastinations and akrasia will not hamper his or her decision.

Last, a critical trait featuring HO is to be *perfectly informed* on, first, the options available and, second, the consequences relevant for the utility associated with each option. For instance, let us imagine that communication for which the lottery prize is paid in annuities rather than in a lump-sum payment is released. If such information is relevant in terms of the utility enjoyed, the model assumes that he or she is certainly aware of it. In sum, if the information is available and relevant in terms of utility, the agent knows it.

The "behavioural revolution" consisted in the endeavor of enriching and eventually revising the HO model in the belief that empirically and psychologically informed assumptions lead to better predictions of human behaviours.

The seminal work by Amos Tversky and Daniel Kahneman showed the merit of such an approach first, followed by the research carried out, among others, by Richard Thaler, one of the fathers of behavioral economics (Thaler 2016).

The "heuristic and bias" program sprang from the behavioural revolution and casts light on several systematic, and consequently predictable, deviations of humans from HO (Khaneman 2011).

Thaler and Sunstein have brilliantly considered such deviations *opportunities* rather than impediments. Since flesh-and-blood humans are hopelessly driven by cognitive biases, why not take advantage of this and *nudge* decision-makers toward desirable behaviors?

Indeed, Sunstein and Thaler consider a 'nudge' to be "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. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates" (Thaler and Sunstein 2008, p 6).

In other words, nudges are kind of interventions that would be irrelevant if the world were populated by *homines oeconomici*, namely agents whose decisions are affected exclusively by bans, coercion, economic incentives, both positive and negative, and necessarily unvailable information up to that time. In other words, nudges can be conceived as interventions leverage on what Thaler calls *supposedly irrelevant factors*, namely factors that it would be correct to suppose to be irrelevant if humans behaved as *homines oeconomici* do.

Such definition of nudges has been deemed too vague, resulting in the difficulty of discerning between a policy that should be considered a nudge and more conventional policies (Hansen 2016, Barton & Grüne-Yanoff 2015).

In the remaining part of the paper, I adopt the term 'nudge' slightly more narrowly than Thaler and Sunstein, following the influental considerations made by Hausman and Welch (2010).
Given the original definition, even reiterating already available information relevant to the options' utilities should be considered a kind of nudge. Indeed, as seen, HO is assumed to be perfectly informed, so reiterating certain information is a supposedly irrelevant factor.
However, to make salient information seems to be a traditional and well-established policymaking strategy, surely not as unconventional as nudges are described. Borrowing the words by Hausman and Welch: "Thaler and Sunstein's characterization of paternalism mistakenly counts giving advice and rational persuasion that aims at the good ofthe advisee as paternalistic" (2010, p. 127). In order to formulate a definition of nudges that account for their disruptive originality, hereinafter I refer to nudges as policy tools leveraging on factors retained to be irrelevant (for HO), *except for reiterating information*.[2]
Now that we have a definition of nudges that pay tribute to their originality, in the next section I will discuss the main ethical issue in nudging decision-makers, namely the lack of transparency.

## II. Transparency in nudging

The definition just outlined is narrow enough to exclude iterating information as nudges but appreciably large enough to include *all* kinds of interventions conceived to leverage human cognitive biases.
Taking into account human cognitive biases in policymaking is twofold. First, it means exploiting cognitive biases, for instance, the default effect. Briefly, the default effect affects our propensity to choose a particular option within a defined set. Basically, the default effect captures the fact that agents will choose a certain option with more probability if it is the option they end up with if they do nothing. For instance, let us consider a topical subject: vaccine choice. Policymakers can arrange the choice environment relevant for the vaccine choice in at least two ways. On the one hand, policymakers can ask citizens to make an appointment proactively (opt-in option). On the other hand, vaccine appointments could be set by default; hence citizens who are not interested in vaccinating are asked to opt-out actively and cancel the vaccine appointment. This last condition turned out to promote a higher number of appointments and, in turn, a higher vaccination rate than the alternative condition (Chapman *et al.*; Lehmann *et al.* 2016).
However, *exploiting* cognitive biases is only one of the two feasible paths. Indeed, policymakers can also shape the choice environment to refrain, or at least mitigate, the

[2] This does not amount to saying that informing cannot be a form of nudging; instead, informing should be considered a form of nudging if cognitive biases are exploited (see Loewenstein & Charter 2017).

effect of cognitive biases, encouraging more careful considerations. An instance of such kind of intervention is providing cooling-off periods when decision-makers face choices that have formerly involved a great deal of regret among peers.

These different typologies of nudges are usually framed referring to the dual-system theory of mind developed by Kahneman and Tversky. Briefly, such theory brings into play two fictional characters, System 1 and System 2, that work in parallel to evaluate the option available and lead to a behaviour. System 1 includes the intuitive, effortless and automatic cognitive processes. Cognitive biases are precisely due to the misuse of system 1, that is, both cases in which we rely on system 1, the automatic pilot, when system 2 should be called into play instead or cases in which system 1-processes mislead our deliberations, as in the case of calculation of probabilities . Instead, system 2 includes all the deliberative, high-level and conscious cognitive processes.

In light of Kahneman and Tversky's theory of mind, we could distinguish between system-1-nudges, namely nudges that leverage cognitive biases, and system-2-nudges, that is nudges that encourage the deliberative process to resist cognitive biases' influences. Such categorization is not just helpful in acknowledging the eterogenity of the nudge theory; rather, it guides us in identifying the exact cases of nudging ethically controversial in liberal democracies.

When nudgees find themselves in a choice environment featured by a system-2-nudge, they can easily recognize the attempt to influence their decisions made by the nudger. In the case of cooling-off periods, for instance, nudgees can easily detect the presence of the policy intervention and arguably become aware of its behavioral aim.

Such kind of transparency on the policymakers' influence attempt is pivotal in liberal democracies where decisional autonomy is an essential value (see Smith et al. 2013; Wilkinson 2012). Evidently, decisions made by agents who live in liberal democracies can, under certain circumstances, be influenced and directed by policymakers, as in the case of bans, coercion and economic incentives. However, what is instead impermissible to policymakers is imposing an influence without citizens being able to detect it, its behavioral aim and, eventually, being able to resist it (Schmidt 2017).

Unfortunately, system-1-nudges could involve this kind of concealed influence. As seen, system-1-nudges leverage human cognitive biases, which are deeply wired into our brain, making their influences typically go unnoticed. Let us consider nudges based on the default effect, specifically the case concerning vaccine appointments just considered. Here, even if nudgees, in fact, could in some way be able to detect the intervention implemented (i.e. the setting of vaccine appointments as default) and eventually its behavioral aim, they would hardly be able to fully recognize the influence exerted by the intervention on their decisions, making it virtually impossible to resist it.

This does not amount to saying that concealed influence attempts *feature all system-1-nudges*. For instance, it is not the case with the fake flies in the urinals adopted in airports worldwide. Here, although automatic processes are harnessed to reduce men's "spillage", nudgees can detect the intervention, the behavioral aim pursued through it, the influence exerted (take the shot!) and, eventually, be able to stupidly resist the influence.

Nevertheless, many system-1-nudges other than nudges based on the default effect seems to easily impose a concealed influence; among them nudges that rest on the framing effect (Tversky & Kahneman, 1981) and the decoy effect (Huber *et al.* 1982).
For this reason, many scholars argue for providing some kind of transparency when such kinds of system-1-nudges are introduced. That is, scholars argued for additional interventions to make it easier for citizens recognizing the influence attempts made by nudgers. We could say that such kinds of intervention are meant to turn some system-1-nudges from tools of covert influence into *tools of transparent influence*.
It is immaterial for the purpose of the paper to exhaustively overview the several proposals advanced to bring concealed influences into light to make ethically justified the employing of challenging system-1-nudges in liberal democracies (Ivanković & Engelen 2019). Here, it shall be sufficient to discuss the two lines along which such proposals have been developed. First, proposals differ from each other in the partition between nudgee and nudgers of the burden required to make nudges' influences actually transparent. If on the one hand, the empirical research on the impact of transparency on nudges' effectiveness provides for cases in which nudgers are asked to disclose information meant to avoid concealed influence (Bruns *et al.*, 2018; Casal *et al.*, 2019; Loewenstein *et al.*, 2015), on the other hand, Bovens (2009) and Ivanković & Engelen (2019), albeit through different modalities, ask for a greater accountability of nudgees, requiring them to be watchful.
Secondly, the debate on what, in fact, should be made transparent to factually defuse the exploiting of hidden influences is not settled. Empirical works on nudges' transparency consider a wide range of information meant to make evident several aspects of nudges, beginning with information concerning the mere existence of an intervention and its behavioral aim. Other than that, information on the effect exploited, the cognitive mechanisms underlying the effect (on which we often know little, see Grüne-Yanoff 2016), the nudge's political aim, and the side effects involved (that is, make salient the fact that nudges could steer some nudgeers toward a behaviour undesirable for them) have been considered.
This section made evident that "transparency" is at the heart of the discussion on the ethics of nudges, other than establishing that the available strategies to pratically make transparent nudges are manifold. This makes it somewhat surprising that, to my best knowledge, nudges' transparency did not duly enter the debate around the ethics of robot-nudging yet, that is, cases in which robots step into the shoes of nudgers.
In the next section, I will briefly overview the current debate on robot-nudging' ethics and set the stage for delving into the debate on transparency in robot-nudging, which, as we will see in section IV, raises questions specific to robot-human interactions.

## III. The ethics of robot-nudging (so far)

The technology to build robots able to influence users' behaviour through nudges (henceforth RN, which stands for robot nudgers) is already available. Factually, RN have already been conceived. For instance, Ali Mehenni and colleagues (2021) experimented the use of nudger dialogue systems as Pepper, a social robot, among children from five to ten years old. Hang and colleagues (2021) investigated if the positive effect on altruism that nudges showed in human-human interaction characterises also cases in which social robots are nudgers.

In addition, it is not that hard to imagine reprogramming robots already built and eventually placed on the market, to make them able to nudge. For example, let us consider the robot trainer developed by Rea and colleagues (2021) to assess the strength of polite and impolite verbal encouragements in steering users to exercise better. This robot could be reprogrammed to convey, other than im(polite) encouragements, sentences designed to prompt peer pressure (see Cialdini & Goldstein 2004), so sentences meant to that make salient that the members of the relevant social network, let us say elderly people, train harder than the user.

Notwithstanding RN are an already available computing technology and foreseeably a widespread one in the near future, the debate on the ethics of robot-nudging is scarce, though not absent.

In 2017, the IEEE Standards Association established the *Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems* committee, and, in 2020, the team "Affective and social dimensions in the spoken interactions" led by Laurence Devillers, along with other colleagues, launched the "Bad Nudge - Bad Robot?" program. The program aims to delve into the risk posed by nudges to vulnerable people.

The current literature on the ethics of RN stresses how when robots are considered as nudgers, new ethical questions emerge compared to those relevant when human-human interactions are in place.

Investigation on robot-nudging's ethics can be roughly divided into two macro areas. The first macro area is devoted to ethical concerns linked to the *behavioral goals* RN should help to achieve. The second macro area instead addresses the ethical questions raised when the *nudging processes* are considered.

Research investigation carried out by Borenstein and Arkin on nudging social justice (2015), by Klincewicz (2019) on stoic ethics, and by Howard and Sparrow (2021) on nudging sexual behaviours belong to the first research area.

Borenstein and Arkin (2015) should be mentioned as well as research belonging to the second macro area. Their paper indeed discussed as well the level and the kind of control on nudges' behavioral aims that should granted to nudgees. Rodogno (2020) points out how social robots, as opposed to human nudgers, could promote behavioral changes influencing users' overall cognitive and affective states. Finally, Calboli and colleagues (2022) discussed the impact of robots' design traits on their strength in nudging.

Within the research investigation on nudging processes in robot-nudging, the debate on transparency completely lacks albeit transparency on the influence attempt made by human nudgers is at the heart of the literature on the ethics of nudging.
In the light of foregoing, it looks fundamental to investigate transparency as an ethical condition for robot-nudging, especially due to the peculiarity of human-robot interactions.
It is worthing to stress that here transparency has as its object the influence imposed on behaviours; hence it is dissimilar to the transparency relevant in the debate on AI explainability. in explainable AI, the focus is instead on users' chance to scrutinize the decisions performed by AI systems.
It is even more earnest focusing on transparency if we consider the fact that, being robots embodied and physically present, they are able to nudge humans both directly and indirectly. Nudging directly means to exploit the physical presence, as in the case of the robot trainer above-mentioned. On the other hand, robots can also nudge indirectly, that is intervene on the choice environment inhabited by the would-be nudgees, as in the case in which a robot rearrange the pantry following behavioral sciences' insights. This last case marks a sharp difference with virtual AI agents.
The following section is meant to pave the way to carry out research to fill this gap and, in turn, begin to see whether the request for transparency in robot-nudging implies ethical issues that do not emerge when human-human interactions are on focus.

## IV. Transparency in robot-nudging

This section aims to provide starting points to integrate issues related to transparency within the debate on the ethics of robot-nudging. In order to do so, I present two broad, ethically relevant points.
First, when human-human interactions are in place, nudgers typically implement nudges to modify *univocally* the choice environment. That is, nudgers do not tailor the nudge's influence to the specific nudgee. Considering the case of vaccine appointments by default, the nudge applies to *all* citizens, regardless of personal tendencies, for instance, in terms of vaccine hesitancy (see SAGE 2014).
In truth, it is worth noticing that customized nudging is available and, in fact, increasingly implemented even when human-human interactions are in place. However, when humans are nudgers, nudges are typically tailored to sub-groups rather than individuals, as in the work by Page and colleagues on reminder text messages to apply for receiving federal student aid. These reminders were customized as meaning that they varied according to the stage in which students were: application not yet started, halfway through, or finished but follow-up requirements were coming (Page *et al.* 2020).
With RN is another matter. For instance, thanks to facial recognition technologies, robots are tools able to customize their actions in accordance with the specific user. Let us recall the case of the robot-trainer developed by Rea and colleagues (cf. section 3). Being fundamental to mentioning the actual reference network to successfully nudge (see

Bicchieri 2014), the robot should convey peer-pressure-triggering information referred to older adults when an older adult uses the robot, and teenagers when a teenager trains. The degree of detailing in identifying the exact reference network can be virtually customized to an individual level. RN can for instance be programmed in a way in which facial recognition technologies are employed to nudge a specific user exclusively among many present or nudge several nudgees differently according to the specific users.

Moreover, this very same strength opens the possibility to customize not only the nudging but also the kind and degree of transparency granted to the nudgee. That is, considering robots customized transparency is available. Hence, the following questions raise: *should it be done? If so, how? Are some customizations improper?*

In other words, the nudgee who knows the robot's ability to nudge can be asked to choose the kind and level of transparency she wants to be in place, should it be deliberately done? Should RN be programmed in such a way? Is so, the nudgee might opt for a different kind of transparency compared to the one considered to be the more suitable by the nudger. For instance, the nudgee could prefer a version of transparency for which the burden to detect and comprehende nudge are totally on her and, contrariwise, the nudger could retain more suitable to take charge of that burden, at least partialy. Should nudgees be enabled to customize transparency?

If so, a second ethically relevant question, strictly connected with this one, emerges: *are there options that should be forbidden being deemed ethically improper?*

Let us consider a case in which an obese person relies upon the help of a robot-nudger to lose weight and virtually save her life. Let us consider a case in which that person is persuaded, rightly or wrongly, that any form of transparency would impede her from reaching the aim and consequently opt for a complete and irrevocable opacity of nudges. Should this - namely, a choice that reminds the Millian case of self-enslavement - be permitted? On the opposite side of the spectrum, there could be cases where nudgees prefer instead, for whatever reason, that the burden required to detect nudges' influence is totally on the nudger's shoulders and that a full range of information should be released by the robot. It could be well the case, although it should be empirically investigated, that such kind of maximum transparency would impair the relationship between nudgees and RN, turning it into a series of annoying exchanges, breaking the harmony of the interaction and as a result hampering nudging processes.[3]

The second ethically relevant question that emerges specifically when human-robot interactions are at hand concerns robots' potential to enhance their ability to nudge.

RN can indeed be programmed to collect data in order to *profile* the nudgee, and so being able to identify the best mix between nudges and transparency in terms of nudges' efficacy. For instance, RN could identify the best timing to both nudge and make transparent the exerted influence in light of a developed model of users' circadian rhythms (Park *et al.* 2010, this case has been considered by Borenstein and Arkin (2016) albeit in a different context). Hence, a second question emerges: *should RN be programmed to collect data to fulfill such an aim?*

[3] It is reasonable to believe that the same could result from boosting strategies, in which decisionmakers are put in the condition to exercise their agency (Grüne-Yanoff & Hertwig 2016).

At first sight, once assumed that privacy issues can be overcome, collecting such data seems an unmissable opportunity. Indeed, these data would be able to help nudgees to independently achieve their behavioral goals by identifying the choice environments helpful to do so. Secondly, the data collected would make able RN suggest nudgees on how to proactively shape the choice environments they are responsible for to make them more likely to achieve the behavioral goals they yearned.
Nevertheless, a thorough analysis reveals the possibility that letting RN collect data on the best mix between transparency and nudge processes could result in severe side effects. The feeling to be observed, and monitored could indeed easily result in psychological reactance, namely the "unpleasant motivational arousal that emerges when people experience a threat to or loss of their free behaviors [...; this] results in behavioral and cognitive efforts to reestablish one's freedom" (Steindl *et al.* 2015, p 205). Unfortunately, psychological reactance is the primary concern among scholars regarding nudges' transparency. If this happens, it would jeopardize the harmony of the interactions between human nudgees and RN, compromise their social interaction, lead nudgees to avoid RN and ultimately curb the chance that users are factually nudged.
The points I just made should be reasonably expected to be just two of the many instances of the ethically relevant questions that transparency in robot-nudging would raise. Even though these ethical issues should be high on the agenda of roboticists and ethicists, factually, they are not. Hopefully, the present work will encourage taking steps in this direction and inspire scholars to explore systematically transparency in robot-nudging.
In the next section, I will summarize the paper's major points and clarify the methodological approach so far implicitly assumed.

## V. Conclusion

In this paper, I discussed the conceptual background of the nudge theory and proposed a working definition of nudges capables of accounting for their originality and status as unconventional policy tools. Then, I engaged with the debate on the ethics of nudges and analyzed the request for transparency. Afterwards, I reviewed the current literature on the ethics of RN, and I stressed how the issues linked to nudges' transparency are overlooked and surprisingly so. In the last section, I discussed two ethically relevant points concerning transparency in robot-nudging in the hope of encouraging further research on the topic.
I conclude the paper by making explicit the methodological assumption underlying the research investigations on transparency in robot-nudging I sketched.
The research line described here would be fruitful developed following the method paradigm called "integrative social robotics" (see Seibt 2016). Integrative social robotics advocates for an interdisciplinary approach, claiming that investigations on what social robots *can* do should advance hand in hand with investigations on what social robots

*should* do (Seibt 2016). This method paradigm aims at the setting of a complex investigation where value-theoretic research is involved since the early stages of social robots' development, being "interactions" what roboticists actual design and so products that inherently imply ethical norms (see the integrative social robotics's quality principles in Seibt *et al.* 2018).

The need to investigate the role that transparency plays in interactions among human nudgeers and RN can be successfully fulfilled following integrative social robotics. This approach enables us to identify and investigate the ethical significance of particular interactions, and doing so concerning transparency in robot-nudging is of paramount importance.

## References


Ali, Mehenni H., Kobylyanskaya, S., Vasilescu, I., Devillers, L. (2021) Nudges with Conversational Agents and Social Robots: A First Experiment with Children at a Primary School. In: D'Haro, L.F., Callejas, Z., Nakamura, S. (eds) *Conversational Dialogue Systems for the Next Decade. Lecture Notes in Electrical Engineering*, vol 704. Springer, Singapore. https://doi.org/10.1007/978-981-15-8395-7_19

Barton A., Grüne-Yanoff, T. (2015) From Libertarian Paternalism to Nudging— and Beyond, *Review of Philosophy and Psychology*, 6(3): 341–59, doi.org/10.1007/s13164-015-0268-x

Bicchieri, C. (2014). *Norms in the Wild: How to Diagnose, Measure and Change Social Norms.* Cambridge: Cambridge University Press.

Borenstein, J., Arkin, R.C. (2015) Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being, *Science and Engineering Ethics, 22(1)*: 31–46. https://doi.org/10.1007/s11948-015-9636-2

Borenstein, J., Arkin, R.C. (2016). Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior, *AI & Society, 32(4)*: 499–507. https://doi.org/10.1007/s00146-016-0684-1

Bovens, L. (2009) The ethics of nudge, in Grüne Yanoff, T. and Hansson, S.O. (eds), *Preference Change: Approaches from Philosophy, Economics and Psychology* , Berlin and New York: Springer: 207 219.

Bruns, H., Kantorowicz Reznichenko, E., Jonsson, M. L. and Rahali, B. (2018) Can nudges be transparent and yet effective? effective? *Journal of Economic Psychology*, 65: 41 59.

Calboli, S., Graziani, P., and Even, J. The Ethics of Robot-nudgers' Design, Proceeding of Robophilosophy 2022 (forthcoming)

Casal, S., Guala, F. and Mittone, L. (2019) On the Transparency of Nudges: An Experiment, *CEEL Working Papers n. 1902*

Chapman, G.B., Li, M., Colby, H., and Yoon, H. (2010) Opting In vs Opting Out of Influenza Vaccination, *JAMA*, 304: 43-44, doi: 10.1001/jama.2010.892.

Goldstein, N.J., Cialdini, R.B., and Griskevicius, V. (2008) A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research* , 35 (3): 472 482. doi.org/10.1086/586910

Grüne-Yanoff, T. (2016), Why Behavioural Policy Needs Mechanistic Evidence, *Economics and Philosophy*, 32(3): 463–83, https://doi.org/10.1017/s0266267115000425

Grüne-Yanoff, T., Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, 26: 149–83.

Hang, C.,Ono T. , Yamada, S. (2021). Designing Nudge Agents that Promote Human Altruism. 10.1007/978-3-030-90525-5_32.

Hansen, P.G. (2016) The Definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove? *European Journal of Risk Regulation*, 7(1): 155–74, doi.org/10.1017/s1867299x00005468

Hansen, P.G., Jespersen, A.M. (2013) Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy, *European Journal of Risk Regulation* 4, 1: 3–28.

Hausman, D. M., B. Welch (2010) Debate: To Nudge or Not to Nudge, *Journal of Political Philosophy* 18(1): 123–36.

Howard, M., Sparrow, R. (2021) Nudge Nudge, Wink Wink: Sex Robots as Social Influencers. In R. Fan, & M.J. Cherry (Eds.) *Sex Robots. Philosophical Studies in Contemporary Culture*, 28. Cham: Springer.

Huber, J., Payne, J.W., & Puto, C. (1982) Adding asymmetrically dominated alternatives: violations of regularity and the similarity hypothesis, *The Journal of Consumer Research*, 9(1): 90–98.

Ivanković, V., Engelen, B. (2019) Nudging, Transparency, and Watchfulness, *Social theory and practice*, 45: 43-73. 10.5840/soctheorpract20191751.

Kahneman, D. (2011) *Thinking, Fast and Slow*. New York: Macmillan.

Klincewicz, M. (2019). Robotic Nudges for Moral Improvement through Stoic Practice, *Techné: Research in Philosophy and Technology*, *23(3)*: 425–455. https://doi.org/10.5840/techne2019122109

Lehmann, B.A., Chapman, G.B., Franssen, F.M., Kok, G., and Ruiter, R.A. (2016) Changing the Default to Promote Influenza Vaccination Among Health Care Workers, *Vaccine*, 34: 1389-92, doi: 10.1016/j.vaccine.2016.01.046.

Levine, D.K. (2020) Is Behavioral Economics Doomed?: The Ordinary versus the Extraordinary. Cambridge, United Kingdom: Open Book Publishers

Loewenstein, G., Bryce, C., Hagmann, D., Rajpal, S. (2015), Warning: You are about to be nudged, *Behavioral Science & Policy*, *1*: 35-42.

Loewenstein, G., & Charter, N. (2017) Putting nudges in perspective. *Behavioural Public Policy*, *1*(1): 26–53. https://doi.org/10.1017/bpp.2016.7

OECD (2017) Behavioural Insights and Public Policy: Lessons from Around the World. Paris: OECD Publishing.

Page L.C., Castleman B.L., Meyer K. (2020) Customized Nudging to Improve FAFSA Completion and Income Verification, *Educational Evaluation and Policy Analysis*, 42(1): 3-21. doi:10.3102/0162373719876916

Park S., Moshkina L., Arkin R.C. (2010) Mood as an affective component for robotic behavior with continuous adaptation via learning momentum, *Proceedings of 10th IEEE-RAS international conference on humanoid robots* (Humanoids 2010).

Rea S., Schneider S., Kanda T. (2021) Is this all you can do? Harder!": The Effects of (Im)Polite Robot Encouragement on Exercise Effort. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*: 225–233, doi:10.1145/3434073.3444660

Rodogno R. (2020) Nudging by Social Robots, In M. Nørskov, J. Seibt, & O.S. Quick (Eds.), *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy*, 337-345. Amsterdam: IOS Press.

Schmidt, A. (2017). "The Power to Nudge." *American Political Science Review 111(2)*, 404-417.
Seibt J. (2016) Integrative Social Robotics—A new method paradigm to solve the description problem and the regulation problem?, in J. Seibt, M. Nørskov, S.S. Andersen, *What Social Robots Can and Should Do*. Springer, New York, 104-114.

Seibt, J., Damholdt, M., Vestergaard, C. (2018) Five Principles of Integrative Social Robotics. 10.3233/978-1-61499-931-7-28.

Smith, C., Goldstein, D., Johnson, E. (2013) Choice without awareness: Ethical and policy implications of defaults. *Journal of Public Policy & Marketing*, 32: 159-172.

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., Greenberg, J. (2015) Understanding Psychological Reactance, *Zeitschrift Für Psychologie*, *223(4)*: 205–14. https://doi.org/10.1027/2151-2604/a000222

SAGE, Strategic Advisory Group of Experts on Immunization (2014) Report of the SAGE Working Group on Vaccine Hesitancy, www.who.int/immunization/sage/meetings/2014/october/SAGE_working_group_revised_report_vaccine_hesitancy.pdf?ua=1.

Thaler, R.H. (2016) *Misbehaving: The Making of Behavioral Economics* (Reprint ed.). W. W. Norton & Company.

Thaler, R.H., Sunstein, C.R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press.

Thaler, R.H., Sunstein, C.R. (2021) *Nudge: The Final Edition* (Revised ed.). Penguin Books.

Tversky, A., Kahneman, D. (1981) Framing of decisions and the psychology of choice, *Science*, 211(4481), pp. 453–8.

Wilkinson, T. M. (2012). Nudging and Manipulation, *Political Studies*, 61(2): 341–355. https://doi.org/10.1111/j.1467-9248.2012.00974.x