Good Intention, Bad Intention, and Algorithm: Rethinking the Value of Nudge in the Era of Artificial Intelligence

The algorithm not only amplifies every detail of human society but also has the same function as the famous nudge technique, i.e. choice architecture, which pushes people toward a certain direction while assuming it's made by their own will. By this nudge-like function of the algorithm, I want to reevaluate the long-controversial issue of the concept of nudge: is this nudge technique harmless? And if it isn't, can we still use this nudge technique even with good intention? I'll start by introducing the concepts of nudge and sludge then talk about their main issues. Third, I'll use three algorithmic examples to demonstrate the consequences of this nudge technique. Fourth, I will address the nature of the nudge technique and the meaning of intention in nudge. Fifth, I'll push the discussion further for an important philosophical issue: the white lie. Finally, I'll summarize my argument and conclude this paper.

Keywords: *Nudge, artificial intelligence, choice architecture, intention, ethics, algorithm*

Acknowledgment

This study was carried out at the Information Law Center of the Institutum Iurisprudentiae of Academia Sinica, was funded by the "Data Safety and Talent Cultivation Project" from the Research Center for Information Technology Innovation of Academia Sinica, and underwent minor revision at The Second Research Division of the Chung-Hua Institution for Economic Research.

I would like to express my deepest gratitude to Professor Wen-Tsong Chiou for his insightful comments on the earlier drafts of this paper. I am also extremely grateful to the reviewers for their valuable comments on this paper. Last but definitely not least, I want to extend my appreciation to Academia Sinica and Chung-Hua Institution for Economic Research for their full support.

Author Information

Chang-Yun Ku, Academia Sinica; Chung-Hua Institution for Economic Research
https://orcid.org/0000-0001-9623-4028

Ku, Chang-Yun. "Good Intention, Bad Intention, and Algorithm: Rethinking the Value of Nudge in the Era of Artificial Intelligence".
Információs Társadalom XXII, no. 4 (2022): 77–85.
https://dx.doi.org/10.22503/inftars.XXII.2022.4.6

1. Introduction

The concept of nudge can be decomposed into two parts: the harmless technique that uses the cognitive limitation of human nature to push people toward a certain direction silently; and the good intention to use this nudge technique to help people make better choices. Thus, this concept is not without controversy: is using this kind of predesigned choice architecture to move people in a certain direction ethical, even with good intention?

Surprisingly, a similar function of the nudge technique also appears in the era of artificial intelligence (AI): the algorithm can affect people without their noticing, can steer people in a certain direction, and even can make people feel that they are making the decision of their own will. As the algorithm is famous for amplifying almost every detail of human society, it also amplifies the consequence of the nudge-like function, which gives us a chance to review the whole concept of nudge from a fresh but solid perspective.

In order to review the entire concept of nudge, i.e. both the harmless nudge technique and the good intention involved in using it, I'll introduce the concept of nudge and its related concept of sludge, as the starting point.

2. The Concepts of Nudge and Sludge

Based on cognitive psychology, Nobel Prize Winner Professor Richard H. Thaler and Professor Cass R. Sunstein (2008) propose a "nudge technique" that they subsequently divided into two subconcepts of "nudge" and "sludge" (Thaler 2018; Sunstein 2020).

As a technique, a nudge is defined as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options of significantly changing their economic incentives" (Thaler and Sunstein 2008, 6). Thus, the nudge technique is not a mandate, it's an intervention that can be easily avoided; and they describe it as "push mildly or poke gently in the ribs … with the elbow"(Thaler and Sunstein 2008, 4). In other words, the nudge technique has two main characteristics: first, it can make people move in a certain direction, i.e. the function of choice architecture; and second, people are not forced to make such decisions.

Thaler and Sunstein believe that the Authority, who has more knowledge and understanding of certain complex issues, should design choice architecture and make people choose certain predictable decisions, i.e. the better one, as made by their own will. And this brings out the second component of the concept of nudge: the intention.

By the intentional consequence to the individual, it can be further divided into two subsets of the nudge technique: the first is the concept of nudge, i.e. a concept that can help the individual make a better choice; and the second is the concept of sludge, i.e. a concept that has the same choice architecture except that it makes a worse choice for the individual. In this sense, according to Sunstein, the sludge is the negative and unpleasant friction "to make a better choice for people" (Sunstein

2020, 7). As Sunstein illustrates in his paper "Sludge Audits," we can tell the difference by whether the consequences are beneficial for an individual's wealth: nudge is for good and sludge is for evil (Sunstein 2020, 6).

The rationale behind the concept of nudge is libertarian paternalism, as Thaler and Sunstein (2008) emphasize. They believe that the motivation of nudge is based on the good intention of the Authority, who has the power not only to help people make a better choice than their own but also to give people the freedom to choose, as it is not a mandate or forced decision.

Thus, if we break the concept of nudge down thoroughly, it's gone even further than we thought, i.e. first it includes the good intention of the nudger; second, choice architecture that pushes the nudgee toward a certain direction not only does not force the nudgee to choose that certain decision, but can also make the nudgee believe that they are making that certain choice by their own free will.

3. Three Main Issues of the Concept of Nudge

Following from the previous section, there are a few issues we need to address if we want to use nudge as an morally legitimate technique to help others. It includes three main issues from different perspectives: the autonomy of the nudgee, the invisible choice architecture, and the intention and intended consequence of the nudger.

The influence on the nudgee's autonomy is the most criticized part that has been brought up when discussing the concept of nudge (Alfano, Carter and Cheong 2018, 301–304; Puaschunder 2018). This issue includes the freedom of the nudgee, the consent of the nudgee, and all other basic rights of the nudgee that are influenced by the nudger. From the perspective of the nudgee's autonomy, the controversy is whether any decision that results from prelimited or preset choice options can present the true expression of the nudgee, even though the decision is made by the nudgee (Schmidt and Engelen 2020).

The second perspective is the invisible choice architecture, and the transparency of it is the main concern of this topic (Möhlmann 2021; Lembcke et al. 2019). The invisible choice architecture puts all kinds of options in front of the nudgee; however, due to its design being based on human cognitive limitation, the function of choice architecture can successfully push the nudgee to choose a certain option as the nudger's expectation, without nudgee's knowing that the decision has been calculated by the nudger (Guihot, Matthew and Suzor 2020).

And last but not least is the perspective that is related to the intentions and the intended consequence of the nudger. The intended consequence is included in the nudger's intention, because if the nudger's intention is bad then it seems pointless to discuss whether using the concept of nudge is morally legitimate. And if the intention is good, it will definitely include the wish to help the nudgee and the belief that the consequence of the nudge will benefit the nudgee. The nudger's intentional consequence here is in the sense of a wishful benefit, rather than an actual one.

However, as Thaler and Sunstein use the subconcepts of nudge and sludge, and claim that "there is no such thing as a neutral design" (Thaler and Sunstein 2008,

3), we can briefly summarize what we have learned in the contexts of nudge and sludge. First, per their claims, it seems that choice architecture is a neutral or harmless tool, even though the fundamental principle is based on human cognitive limitation; second, it looks like good intention can outweigh the usage of this predesigned choice architecture, and justify its influence on the basic rights of the nudgee.

The descriptions above somehow require our further examination. Is this predesigned choice architecture, which builds on the foundation of human cognitive limitation, really harmless or neutral? And is the nudger's good intention enough for us to justify influencing the nudgee's basic rights by using a predesigned choice architecture? In order to answer these questions, I believe we can use the advanced choice architecture, i.e. the algorithm and its effect of Hypernudge, to illustrate my points.

4. The Advanced Choice Architecture: The Algorithm and Hypernudge

If Thaler and Sunstein's elaboration is the basic understanding of the technique of choice architecture and the concept of nudge and sludge, then Professor Karen Yeung's (2017) illustration of the Big Data analysis technology is the advanced and even more powerful version of it: the algorithm and the hypernudge.

According to Yeung's research, Big Data analysis technology is a choice architecture of information that optimizes the personal choice environment by feeding back the data on personal decisions and the algorithmic technology. With timely data feedback and a correlation-finding function, the algorithm "dynamically configures the contexts of the user's informational choice and consequentially affects that user's choice" (Yeung 2017, 6). Because the result of limiting personal choice is obvious, however, the whole process of the algorithmic limitation is too subtle for the individual to notice, thus, Yeung describes the effect of it as the "Hypernudge." According to Yeung, eventually these feedback data will be used for analyze the behavioral trends of the whole population, as the "Surveillance Capitalism" that Professor Shoshana Zuboff describes (Yeung 2017, 15).

The effect of the algorithmic hypernudge (or the "hypersludge") is accurately embodied in the era of AI, as Yeung warns. Algorithmic hypernudge can be divided into three forms, namely, three different methods for pushing the individual to choose a predesigned or preset option.

The first form is the general limitation. This limitation applies to everyone who may want to access certain information. The case of Google Shopping is one example (Picht and Loderer 2019, 408–410). Google used a preset algorithm to list its own Google Shopping website at the top of the first page of its search results; and meanwhile, Google used a series of criteria to demote competitors' websites in the ranking of search results. According to the European Commission, search results that appear on the first page have a 95% click-through rate (CTR) in comparison to the 1% CTR for results appearing on the second page (European Commission 2017). Google thus gave Google Shopping tremendous advantage simply by placing it at the top of the first page of search results.

The second form of hypernudge is homogeneous information feedback. For example, during an election, Meta's (formerly named Facebook) algorithm provides information on certain political parties to certain groups of people, and thus causes a filter bubble effect on those people (Confessore 2018); and this informational predesign ability has already been proved by Mata itself (Kramer, Guillory and Hancock 2014; Verma 2014). In contrast to the unintended bias result from the historical data training (Chiou 2018), the filter bubble effect makes individuals only receive homogeneous information that is similar to their own opinions, by intended preselecting criteria in the information feedback loop. Because of the homogeneous information, the filter bubble effect reinforces individuals' beliefs on a certain topic, and thus made a decision base on it. For those people who are affected, the "reality" they perceive is totally different from the societies outside of them.

The third form is providing personalized information. This kind of personalized information is based on the algorithmic predictions of users' personalities, characters, preferences, etc. (Helbing et al. 2018). For example, Uber uses the algorithm to determine the price based on their algorithmic prediction of the individual's willingness to pay, namely the personalized pricing, which causes the same distance to have a different price for different individuals (Mahdawi 2018). And because the users in this context have no suspicion that they have been treated differently, they are willingly to choose from the options provided by the algorithm, and pay the personalized pricing unknowingly.

As these examples show, algorithmic hypernudge pushes people to make certain choices or to move in a certain direction, without being noticed by those people. The mechanism of hypernudge is that it limits the information that the individual can receive, and thus it limits the choice that the individual can make, and further, it even changes the individual's perceptions, simply by the functions of the algorithm.

5. Rethinking the Value of Nudge in the Algorithmic Era

The form of general limitation in the case of Google Shopping shows that, with human cognitive limitation, people actually have very limited attention, even when they are provided with all of the information. And in combination with choice architecture, people can focus only on those choice options that are predesigned for them to perceive, and unable to be aware of all the options. Thus, in this sense, the claim of "provide all the choice or information" in the context of the nudge, which the nudger being aware of people's cognitive limitation, is actually equal to manipulating people by the predesigned choice architecture.

The form of providing personalized information even emphasizes our above point. When we are provided with certain algorithmic information or options, we hardly assume that "there is more information" that is hidden from us, based on our assumption that the algorithm is neutral. And, for the same reason, it is also impossible for us to imagine that the information we receive will be different from the information that others receive. But these facts in the previous section precisely show the biased nature of the algorithm, and so does the choice architecture. At some

level, it's correspondent to the claim that "there is no such thing as neutral design"; however, it should be rephrased as "there is no such thing as neutral choice architecture." The very concept of choice architecture implies the biased nature of it.

And the form of homogeneous information feedback in the case of Meta somehow answers our question related to the nudger's good intention. The algorithm was designed to help people make faster and better choices; by feeding a user's personal behavioral or preference data, the user can receive information that related directly to them, focus on the theme they care about, and shorten their decision-making time. However, the example of the filter bubble effect shows that being able to help people make decisions more efficiently requires sacrificing their right to receive all of the information. This kind of "helping people by hurting them" isn't morally legitimate at all; even if we put aside that most of the time, the human good intention is not comprehensive, and the wishful consequence of the good intention could go out of our control and cause unexpected or even unforeseen results.

All three of these examples of algorithmic hypernudge point to a clear conclusion: the only function of choice architecture is nothing but to limit the information to the nudgee, in order to achieve the predesigned result of the nudger's expectation.

Although the consequence above is obvious, we might question whether we need to condemn the nudger's action, if the nudgee agrees to be nudged by the nudger. However, the point here is not about the nudgee's agreement to be nudged; it is about the means or the method that the nudger chooses in order to help the nudgee. The nudgee's consent doesn't affect the main purpose of this paper, namely, to evaluate the moral legitimacy of the concept of nudge.

So, let's go back to our earlier questions. First, is choice architecture a neutral or harmless tool? As the consequence above mentions, choice architecture in nudge is by no means a neutral or harmless tool. Building on the knowledge of human cognitive limitation to design the choice architecture, the nudger only exploits nudgee's limited attention and forces nudgee to choose from options within a certain predesigned range.

Second, as to the Archimedean point of the nudge, can the nudger's good intention be the justification for the impact of the individual's basic rights? The answer is no. The essence of nudge is helping people by limiting their basic right to choose without telling them certain information. This claim of "hurting people for their own good" is absurd and immoral, even if the intention is good.

The discussion of the essence of nudge reminds us of a familiar topic in ethics, namely, the issue of the white lie. In the next section, I would like to talk about this centuries-old question in philosophy and highlight the main points of discussion in this paper.

6. One Step Further: Is the White Lie Permissible?

Following revealing the nature of nudge, let's push the discussion further. Morally speaking, when we evaluate the moral value of a behavior, we are generally referring to the whole decision-making process of it. The whole decision-making process

of a behavior includes three parts: the intention of the individual to take that action, the means that the individual chooses in order to achieve the end, and the consequence of that behavior. Thus, theoretically, there is no single part that can represent the whole moral value of the behavior. And, from this point of view, the moral value of the good intention of the concept of nudge isn't enough to justify the choice of morally wrongful means, since the value of the intention and the means are both included in the moral evaluation of the whole concept of nudge.

This illustration of the value of nudge somehow leads us to answer a centuries-old philosophical question: Is the white lie permissible? According to our presumption, the intention, the means, and the consequence of the behavior are all included in the moral evaluation of the behavior. As for the issue of the white lie, the intention of the white lie is without doubt good, the intentional consequence is kindness to the people that will be affected by the truth, and the means is lying to them. However, as we saw in the concept of nudge, the case of the white lie is also not permissible from the point of view of moral evaluation.

Of course, it's reasonable to ask what if the consequences or effects are too small to notice, do we still consider this behavior unethical? For example, if we tell a white lie only to be polite when we don't want to participate in an event; or the nudge only has the minimum effect like Baldwin classifies as "the first-degree nudge", which provides people only with simple information or a reminder (Baldwin 2014, 835). In these cases, do we still consider that the white lie or nudge is unjustifiable? The answer is the same, of course: to tell a white lie or use a minimum nudge is morally impermissible.

Moral value is nonnegotiable, it is and should be the most solid part of the essence of human behavior or human character. Although, admittedly, we often face difficult choices and moral dilemmas in our daily lives, but this neither gives us an excuse to choose the morally wrong path or the wrongful means to achieve our goal, nor blurs the line between what is right and what is wrong.

No matter whether it's in the case of nudge or in the example of the white lie, in order to achieve the human's genuine goal to help others, the freedom for the human to choose a proper method is obvious and crucial, and the options are where the moral value is contained. Although the technique of nudge is effective, or the effect on our human autonomy might be as little as possible, but none of this gives us a reason to believe that the concept of nudge is morally permissible to use, especially when we know that the nature of choice architecture isn't neutral or harmless.

Therefore, concerning the question of whether a white lie is permissible, our answer is the same as whether the nudge technique as a method is morally permissible, i.e. the point isn't about the effect on the individual or the nudgee is significant or not, it's all about not choosing a morally wrong means to achieve our genuine ends; and the answer is: No.

7. Conclusion

Hypernudge in the algorithmic era shows that the meaning of nudge is actually using the method of information limitation to force the individual to choose a certain

option. Choice architecture is by no means a neutral tool; even the nudger has a good intention and seeking to help people, it can't outweigh the fact that it's morally wrong to choose wrongful means to achieve the ends.

It is crucial when it comes to the question of sacrificing people's rights in order to help them toward a better future; the decision of whether to "be helped" should be left to the nudgees, not the nudger. Even with the nudgees' consent, however, the more important question should be why do we need to use this kind of secretive technique of information limitation to achieve our goal.

So, if someone knows the essence of nudge but still wants to use the nudge technique to help people, then we must ask: Why?

References

Alfano, Mark, J. Adam Carter, and Mark Cheong. "Technological Seduction and Self-radicalization." *Journal of the American Philosophical Association* 4, no. 3 (2018): 298–322. https://doi.org/10.1017/apa.2018.27

Baldwin, Robert. "From Regulation to Behaviour Change: Giving Nudge the Third Degree." *Modern Law Review* 77, no. 6 (2014): 831–857. https://doi.org/10.1111/1468-2230.12094

Chiou, Wen-Tsong. "Evolving Issues of Data Protection and the Conundrum of Antidiscrimination in Artificial Intelligence." In *Emerging Legal Challenges in the Era of Artificial Intelligence*, edited by Ching-Yi Liu, 149–175. Taipei: Angle Publishing, 2018. https://idv.sinica.edu.tw/wentsong/pdf/20181101.pdf

Confessore, Nicholas. "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far." *The New York Times*, April 4, 2018. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

European Commission. "Antitrust: Commission Fines Google €2.42 Billion for Abusing Dominance as Search Engine by Giving Illegal Advantage to Own Comparison Shopping Service – Factsheet." June 27, 2017. https://ec.europa.eu/commission/presscorner/detail/en/MEMO_17_1785

Guihot, Michael, Anne F. Matthew, and Nicolas P. Suzor. "Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence." *Vanderbilt Journal of Entertainment and Technology Law* 20, no. 2 (2020): 385–456. https://scholarship.law.vanderbilt.edu/jetlaw/vol20/iss2/2

Helbing, Dirk, Bruno S. Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hanger, Yvonne Hofstetter, Jeroen van den Hoven, Roberto V. Zicari, and Andrej Zwitter. "Will Democracy Survive Big Data and Artificial Intelligence?" In *Towards Digital Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution*, edited by Dirk Helbing, 73–98. Cham: Springer, 2018. https://doi.org/10.1007/978-3-319-90869-4

Kramer, Adam D., Jamie E. Guillory, and Jeffrey T Hancock. "Experimental Evidence of Massive-scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Science* 111, no. 24 (2014): 8788–8790. https://doi.org/10.1073/pnas.1320040111

Lembcke, Tim-Benjamin, Nils Engelbrecht, Alfred Benedikt Brendel, and Lutz Kolbe. "To Nudge or Not to Nudge: Ethical Considerations of Digital Nudging Based on Its Behavioral Economics Roots." In *Proceedings of the 27th European Conference on Information Systems*, Stockholm-Uppsala, Sweden: European Conference on Information Systems 2019. https://aisel.aisnet.org/ecis2019_rp/95

Mahdawi, Arwa. "Is Your Friend Getting a Cheaper Uber Fare Than You Are?" *The Guardian*, April 13, 2018. https://www.theguardian.com/commentisfree/2018/apr/13/uber-lyft-prices-personalized-data

Möhlmann, Mareike. "Algorithmic Nudges Don't Have to Be Unethical." *Harvard Business Review*, April 22, 2021. https://hbr.org/2021/04/algorithmic-nudges-dont-have-to-be-unethical

Picht, Peter Georg, and Gaspare Tazio Loderer. "Framing Algorithms: Competition Law and (Other) Regulatory Tools." *World Competition* 42, no. 3 (2019): 391–417. https://doi.org/10.5167/uzh-181193

Puaschunder, Julia M. "Nudgital: Critique of Behavioral Political Economy." In *Proceedings of the 9th International RAIS Conference on Social Sciences and Humanities*, 54–76. New Jersey: Research Association for Interdisciplinary Studies at The Erdman Center at Princeton University, 2018. http://dx.doi.org/10.2139/ssrn.2926276

Schmidt, Andreas T., and Bart Engelen. "The Ethics of Nudging: An Overview." *Philosophy Compass* 15, no. 4 (2020): e12658. https://doi.org/10.1111/phc3.12658

Sunstein, Cass R. "Sludge Audits." *Behavioural Public Policy* 6, no. 4 (2020): 1–20. https://doi.org/10.1017/bpp.2019.32

Thaler, Richard H. "Nudge, Not Sludge." *Science* 361, no. 6401 (August 3, 2018): 431. https://doi.org/10.1126/science.aau9241

Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth and Happiness*. New Haven, CT: Yale University Press, 2008.

Verma, M. Inder. "Editorial Expression of Concern: Experimental Evidence of Massive-scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences* 111, no. 29 (2014): 10779. https://doi.org/10.1073/pnas.1412583111

Yeung, Karen. "'Hypernudge': Big Data as a Mode of Regulation by Design." *Information, Communication & Society* 20, no. 1 (2017): 118–136. https://doi.org/10.1080/1369118X.2016.1186713