CESifo WORKING PAPERS

9939
2022

September 2022

Meta-Nudging Honesty: Past, Present, and Future of the Research Frontier

Eugen Dimant, Shaul Shalvi

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9939*

Meta-Nudging Honesty: Past, Present, and Future of the Research Frontier

Abstract

Achieving successful and long-lasting behavior change via nudging comes with challenges. This is particularly true when choice architects attempt to change behavior that is collectively harmful but individually beneficial, such as dishonesty. Here, we introduce the concept of 'meta-nudging' and illustrate its potential benefits in the context of promoting honesty. The meta-nudging approach implies that instead of nudging end-users directly, one would nudge them indirectly via "social influencers". That is, one can arguably achieve better success by changing the behavior of those who have the ability to enforce other's behavior and norm adherence. We argue that this represents a promising new behavior change approach that helps overcome some of the challenges that the classical nudging approach has faced. We use the case of nudging honesty to develop the theoretical foundation of meta-nudging and discuss avenues for future work.

JEL-Codes: C910, D010.

Keywords: behaviour change, honesty, lying, nudging.

*Eugen Dimant**
University of Pennsylvania
Philadelphia / PA / USA
edimant@sas.upenn.edu

Shaul Shalvi
University of Amsterdam
Amsterdam / The Netherlands
s.shalvi@uva.nl

*corresponding author

This version: September 12, 2022

The most recent version of the working paper can always be downloaded following this link: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4081493

Forthcoming in: Current Opinion in Psychology.

This work was financially supported by the German Research Foundation (DFG) under Germany's Excellence Strategy — EXC 2126/1– 390838866.

Introduction

Historically, the concept of nudging has been focused on identifying and changing behavior at the *individual* level [1]. While many success stories suggest that nudges can be effective [2], extant evidence also points out that behavior change is difficult, often produces small effect sizes, sometimes fails, and may even backfire. This is especially true when attempting to achieve *long-lasting* behavior change that extends beyond the time window of the intervention [3–7]. In fact, the effectiveness of nudging has been shown to be highly variable and sensitive to the exact context [8–10], thus making it challenging to select the most potent interventions prior to their implementation, which can be a costly trial-and-error loop.[1]

Recently, scholars have urged a reconsideration of the classical nudging approach that focuses on the individual by putting more emphasis on the environment in which these individuals operate. In turn, this should help behavioral science to transition from amending choice *architecture* to creating choice *infrastructure* [16–18].

Here we propose that 'meta-nudging' constitutes one such promising approach. The central idea of this approach is that rather than targeting individual behavior change *directly*, a more promising way is to change behavior *indirectly*: target individuals – the *social influencers* – who are in positions of power and maintain a level of authority that gives them the ability to enforce good behavior of their subordinates [19].

Meta-Nudging Approach Defined

While nudges that directly target individual behavior change have shown success, this classical approach to nudging has also raised concerns in the scientific community. For example, the focus on individual-level solution has been argued to potentially crowd-out systemic changes – the focus on individual-level changes result in less focus being put on system changes –, thus leading behavioral public policy astray [17].

What could the next generation of nudging that meets this premise look like? One such promising new approach has been coined 'meta-nudging' and suggests that one can also successfully nudge individuals *indirectly* by harnessing the power of social norms enforcement [19]. That is, by targeting those who enforce behavior – rather than those whose behavior one wants to alter – behavioral interventions would aim at nudging individuals in positions of power who have the ability to enforce the transgressors' adherence to social norms.

Research by Dimant and Gesche [19] suggests that 'norm-nudging' can be a potent application of the meta-nudging approach. Norm-nudging, which is a special case of behavioral nudging, aims at eliciting and changing existing social norms through systematic variation of social expectations. This approach has been theoretically conceptualized by [20] in that norm-nudge interventions aim at changing either the beliefs about what others in one's

[1]For behavioral public policy to be effective and to have "bite", the underlying evidence that informs the policies needs to be robust. To achieve this, recent trends in the academic community include the use of prediction markets that harness the forecasting ability of individuals to predict the replicability of existing interventions and the effectiveness of future ones [11–13]. This includes the implementation of so-called 'megastudies' in which independent teams of scholars test different interventions to achieve behavior change [14], as well as meta-analytical evaluation of existing research, published and unpublished, to identify impact and robustness of interventions while also accounting for publication bias as much as possible [2, 15].

reference network *do* (descriptive element of the norm, first-order belief) or what others in one's reference network approve others to do (injunctive element of the norm, second-order belief). The effectiveness of norm-nudging results from targeting (at least) one of three aspects: (i) pointing out bad norms that are currently in place, (ii) defining good norms more clearly, and (iii) facilitating the enforcement of good norms [21–24]. Evidence suggests that norm enforcement is generally prevalent [25, 26], particularly so in 'tight' societies [27], and that enforcement behavior can also be successfully nudged via norm-nudges [19].

There are two advantages for meta-nudging over traditional, direct nudging. The first advantage is the underlying incentive system of the nudgee under which the behavioral intervention operates. In the classical nudging approach, the nudgee often engages in behavior that is beneficial at the individual level (such as driving a car), whereas behavior change that benefits the society (for example, riding a bike instead) would mean to incur individual costs (a reduction in convenience) in favor of the collective gain (reduction in CO_2 emissions). Consequently, for a nudge to be effective, the intervention needs to overcome two forces that run counter to the target behavior: individual inertia (or disapproval of the target behavior) and opposing incentives (e.g., foregone pleasure of staying dry when driving a car rather the bike when it is raining). In addition, cognitive dissonance from abandoning one's initial (selfish) behavior is typically present in such instances and further challenges the effectiveness of the nudge. Individuals for whom this cannot be achieved are typically characterized as 'un-nudgeable' [28].

Meta-nudging, on the other hand, targets social influencers who can enforce good norms via social (or financial) pressure which in turn prevents bad norms from spreading. While the meta-nudge also needs to be potent enough to overcome the influencer's inertia and other related individual costs such as the fear of potential retaliation from the subordinate nudgee, there are now also counteracting forces that facilitate the success of the meta-nudge. For example, any utility that the influencer derives from impacting others' behavior or from enforcing norms, which motivates the influencer to positively react to the nudge. Indeed, supporting those assumptions, influencers were found to be 'social trendsetters' who are ready to bear a cost to initiate change because they are usually less sensitive to risk [29].

The second advantage of meta-nudging is that behavioral interventions that rely on delegated policing ("hired guns") might both be perceived less intrusive and more successful in that they would capitalize on existing peer mechanisms [30]. Arguably, this would increase the acceptability of enforcement, which has been shown to be a crucial ingredient of successful norm enforcement [31]. In what follows, we will apply these insights to the case study of nudging honesty and discuss promising avenues for future research.

Meta-Nudging Approach Applied to Dishonesty

Changing behavior in the context of curbing dishonesty is challenging because of diverging incentives: dishonesty is often individually beneficial but collectively harmful. Thus, any behavioral intervention aimed at changing behavior directly needs to convince the individual to forego an individual benefit in favor of the collective good. Evidently, this is not only the case when societal norms about the proper behavior are vague and contain moral wiggle-room [32], but also when norms are firmly established and followed by peers [22, 33, 34].

Take for example, the norm of honest behavior, which is praised and socially desirable. Nonetheless, high-profile and systemic cases of dishonesty still persist (see, e.g., the recent Enron, Madoff, and Volkswagen scandals) [35]. Research on these topics suggests that the effectiveness of reducing dishonesty via nudging varies [8, 36] and can be explained by the various factors that determine dishonest behavior, to which we will turn below.

Most existing research has focused on understanding the mechanisms underlying dishonest behavior, with the premise that gaining such an understanding would allow crafting interventions to increase honesty. The key mechanisms identified include one's ability to exploit moral wiggle-rooms via self-serving justification [37, 38]. That is, individuals are able to abuse an existing moral wiggle-room by reinterpreting, distorting, or purposefully forgetting existing evidence favoring norms of honesty [32, 36, 39]. Another mechanism driving dishonest behavior is people's tendency to purposely select, seek, and process available information, which allows individuals to remain ignorant and maintain plausible deniability [40–42]. This line of research emphasizes dishonesty as largely independent of others [43].

Recent work further demonstrated the large impact one's (dis)honesty has on others (dis)honesty. Specifically, when considering settings in which one finds justification for one's own dishonesty in the dishonesty of peers [44–46], people lie a lot. The core insight from this research is that social reinforcement via observing and being observed by one's peers is interpreted as a signal of the dominant social norm, which can accelerate the contagion of dishonesty [4, 21, 22, 47]. For example, [45] found that in a repeated interaction between two individuals, in which they both stand to benefit from each other's dishonesty, when a group member signals dishonesty on the very first move, such behavior more than doubles the group's overall dishonesty compared with a situation in which no such signal exists. Recent field research indeed confirmed that the likelihood of a call-center employee to be (dis)honest varies as a function of the (dis)honesty of those sitting in their proximity [48]. Taken together, those finding demonstrates the promise in meta-nudging honesty. Given that people one's (dis)honesty has such strong impact on those one interacts with, demonstrates the promise in interventions aimed at meta-nudging honesty.

Conclusion and Future Directions in Meta-Nudging

Sustained behavior change is hard. This is even true when individuals are 'nudgeable' and have a pre-disposition that favors behaviors that one can generally agree on is largely beneficial, such as eating healthier. However, it is arguably even harder to try to change behavior such as dishonesty which – even though it is detrimental on a collective level and potentially also violates existing social norms – is beneficial at the individual level. This is because individual and collective incentives are misaligned and behavioral interventions need to be potent enough to help the individual to put more weight on the latter. As we argue throughout the paper, we believe that the concept of 'meta-nudging' presents a promising new approach to yield more successful behavioral interventions.

More specifically, building upon the meta-analytical insights suggesting a strong impact of one's (dis)honesty on others' (dis)honesty, we can construct different forms of meta-nudging. For example, since the level of dishonesty has been found to be sensitive to financial incentives, nudging influencers to enforce deviance via costly punishment – as

successfully tested in the original meta-nudging approach by [19] – is a promising avenue. Alternatively, since transgressors factor in the negative externalities that their behavior produces, influencers can attempt to highlight those when nudging honesty. Thus far, this approach has been mostly tested successfully in individual-decision environments [19, 49]. Investigating whether these interventions are also successful in collaborative environments that are characterized by social interactions remains an empirical question.

We see this approach as complementary to the classical nudging approach allowing the choice architect to select from a wider array of tools. The correct tool will be context-dependent, will require testing and re-testing, and a careful roll-out when attempting to achieve success at scale [50]. By complementing the arsenal of behavioral change techniques that target individual decision-making (streamlining decision environments, defaults etc.) with the 'meta-nudging' approach, policy-makers can build momentum at the collective level. The long-term success of such an approach remains an empirical question and represents a potent future direction the behavioral science field can head towards.

References

[1] R. H. Thaler and C. R. Sunstein, *Nudge: The final edition*. Penguin, 2021.

[2] S. DellaVigna and E. Linos, "RCTs to scale: Comprehensive evidence from two nudge units," *Econometrica*, vol. 90, no. 1, pp. 81–116, 2022.

[3] J. Beshears, J. J. Choi, D. Laibson, B. C. Madrian, and K. L. Milkman, "The effect of providing peer information on retirement savings decisions," *The Journal of Finance*, vol. 70, no. 3, pp. 1161–1201, 2015.

[4] G. Bolton, E. Dimant, and U. Schmidt, "Observability and social image: On the robustness and fragility of reciprocity," *Journal of Economic Behavior & Organization*, vol. 191, pp. 946–964, 2021.

[5] A. Brandon, P. J. Ferraro, J. A. List, R. D. Metcalfe, M. K. Price, *et al.*, "Do the effects of social nudges persist? theory and evidence from 38 natural field experiments," Working Paper, 2022.

[6] M. Gelfand, R. Li, E. Stamkou, D. Pieper, E. Denison, *et al.*, "Persuading republicans and democrats to comply with mask wearing: An intervention tournament," *Journal of Experimental Social Psychology*, 2022.

[7] C. Morvinski, S. Saccardo, and O. Amir, "Mis-nudging morality," *Management Science*, 2022.

[8] J. Beshears and H. Kosowsky, "Nudging: Progress to date and future directions," *Organizational Behavior and Human Decision Processes*, vol. 161, pp. 3–19, 2020.

[9] E. Dimant, "Hate trumps love: The impact of political polarization on social preferences," Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3680871, 2022.

[10] R. Hertwig and N. Mazar, "Toward a taxonomy and review of honesty interventions," Working Paper, 2022.

[11] C. F. Camerer, A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, *et al.*, "Evaluating the replicability of social science experiments in nature and science between 2010 and 2015," *Nature Human Behaviour*, vol. 2, no. 9, pp. 637–644, 2018.

[12] S. DellaVigna, D. Pope, and E. Vivalt, "Predict science to improve science," *Science*, vol. 366, no. 6464, pp. 428–429, 2019.

[13] E. Dimant, E. G. Clemente, D. Pieper, A. Dreber, and M. J. Gelfand, "Politicizing mask-wearing: Predicting the success of behavioral interventions among republicans and democrats in the u.s.," *Scientific Reports*, 2022.

[14] K. L. Milkman, M. S. Patel, L. Gandhi, H. N. Graci, D. M. Gromet, *et al.*, "A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment," *Proceedings of the National Academy of Sciences*, vol. 118, no. 20, 2021.

[15] N. C. Köbis, B. Verschuere, Y. Bereby-Meyer, D. Rand, and S. Shalvi, "Intuitive honesty versus dishonesty: Meta-analytic evidence," *Perspectives on Psychological Science*, vol. 14, no. 5, pp. 778–796, 2019.

[16] M. Hallsworth and E. Kirkman, *Behavioral insights*. MIT Press, 2020.

[17] N. Chater and G. Loewenstein, "The i-frame and the s-frame: How focusing on the individual-level solutions has led behavioral public policy astray," Working Paper, 2022.

[18] C. R. Sunstein, "The distributional effects of nudges," *Nature Human Behaviour*, vol. 6, no. 1, pp. 9–10, 2022.

[19] E. Dimant and T. Gesche, "Nudging enforcers: How norm perceptions and motives for lying shape sanctions," Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3664995, 2021.

[20] C. Bicchieri and E. Dimant, "Nudging with care: The risks and benefits of social information," *Public Choice*, pp. 1–22, 2019.

[21] E. Dimant, "Contagion of pro-and anti-social behavior among peers and the role of social proximity," *Journal of Economic Psychology*, vol. 73, pp. 66–88, 2019.

[22] C. Bicchieri, E. Dimant, S. Gächter, and D. Nosenzo, "Social proximity and the erosion of norm compliance," *Games and Economic Behavior*, vol. 132, pp. 59–72, 2022.

[23] E. Dimant, "Distributions matter: Measuring the tightness and looseness of social norms," Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.4107802, 2022.

[24] J. A. Yip and M. E. Schweitzer, "Norms for behavioral change (nbc) model: How injunctive norms and enforcement shift descriptive norms in science," *Organizational Behavior and Human Decision Processes*, vol. 168, p. 104 109, 2022.

[25] E. Fehr and S. Gachter, "Cooperation and punishment in public goods experiments," *American Economic Review*, vol. 90, no. 4, pp. 980–994, 2000.

[26] L. Balafoutas and N. Nikiforakis, "Norm enforcement in the city: A natural field experiment," *European Economic Review*, vol. 56, no. 8, pp. 1773–1785, 2012.

[27] M. J. Gelfand, J. L. Raver, L. Nishii, L. M. Leslie, J. Lun, *et al.*, "Differences between tight and loose cultures: A 33-nation study," *science*, vol. 332, no. 6033, pp. 1100–1104, 2011.

[28] D. de Ridder, F. Kroese, and L. van Gestel, "Nudgeability: Mapping conditions of susceptibility to nudge influence," *Perspectives on Psychological Science*, p. 1 745 691 621 995 183, 2021.

[29] C. Bicchieri, *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press, 2016.

[30] J. Andreoni and L. K. Gee, "Gun for hire: Delegated enforcement and peer punishment in public goods provision," *Journal of Public Economics*, vol. 96, no. 11-12, pp. 1036–1046, 2012.

[31] C. Bicchieri, E. Dimant, and E. Xiao, "Deviant or wrong? the effects of norm information on the efficacy of punishment," *Journal of Economic Behavior & Organization*, vol. 188, pp. 209–235, 2021.

[32] C. Bicchieri, E. Dimant, and S. Sonderegger, "It's not a lie if you believe the norm does not apply: Conditional norm-following with strategic beliefs," Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3326146, 2021.

[33] C. Deutscher, E. Dimant, and B. R. Humphreys, "Match fixing and sports betting in football: Empirical evidence from the German Bundesliga," Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.2910662, 2021.

[34] E. Dimant, M. Gelfand, A. Hochleitner, and S. Sonderegger, "Strategic behavior with tight, loose, and polarized norms," Working Paper Available at SSRN: https://bit.ly/3ryY3Pc, 2022.

[35] A. Cohn, E. Fehr, and M. A. Maréchal, "Business culture and dishonesty in the banking industry," *Nature*, vol. 516, no. 7529, pp. 86–89, 2014.

[36] E. Dimant, G. A. Van Kleef, and S. Shalvi, "Requiem for a nudge: Framing effects in nudging honesty," *Journal of Economic Behavior & Organization*, vol. 172, pp. 247–266, 2020.

[37] S. Shalvi, J. Dana, M. J. Handgraaf, and C. K. De Dreu, "Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior," *Organizational Behavior and Human Decision Processes*, vol. 115, no. 2, pp. 181–190, 2011.

[38] S. Shalvi, F. Gino, R. Barkan, and S. Ayal, "Self-serving justifications: Doing wrong and feeling moral," *Current Directions in Psychological Science*, vol. 24, no. 2, pp. 125–130, 2015.

[39] S. Saccardo and M. Serra-Garcia, "Cognitive flexibility or moral commitment? evidence of anticipated belief distortion," Working Paper, 2022.

[40] R. Golman, D. Hagmann, and G. Loewenstein, "Information avoidance," *Journal of Economic Literature*, vol. 55, no. 1, pp. 96–135, 2017.

[41] E. Dimant, F. Galeotti, and M. C. Villeval, "Norm-formation and the role of information acquisition," Mimeo, 2022.

[42] L. Vu, I. Soraperra, M. Leib, J. van der Weele, and S. Shalvi, "Willful ignorance: A meta-analysis," Working Paper, 2022.

[43] N. Mazar and D. Ariely, "Dishonesty in everyday life and its policy implications," *Journal of Public Policy & Marketing*, vol. 25, no. 1, pp. 117–126, 2006.

[44] O. Weisel and S. Shalvi, "The collaborative roots of corruption," *Proceedings of the National Academy of Sciences*, vol. 112, no. 34, pp. 10 651–10 656, 2015.

[45] M. Leib, N. Köbis, I. Soraperra, O. Weisel, and S. Shalvi, "Collaborative dishonesty: A meta-analytic review.," *Psychological Bulletin*, vol. 147, no. 12, p. 1241, 2022.

[46] O. Weisel and S. Shalvi, "Moral currencies: Explaining corrupt collaboration," *Current Opinion in Psychology*, vol. 44, pp. 270–274, 2022.

[47] Z. B. Ren, E. Dimant, and M. E. Schweitzer, "Social motives for sharing conspiracy theories," Working Paper, 2022.

[48] R. Ferrali, "Is honesty or dishonesty more contagious? Evidence from the field," Working Paper, 2020.

[49] J. J. Zlatev, D. P. Daniels, H. Kim, and M. A. Neale, "Default neglect in attempts at social influence," *Proceedings of the National Academy of Sciences*, vol. 114, no. 52, pp. 13 643–13 648, 2017.

[50] J. List, *The Voltage Effect*. Penguin Books Limited, 2022, ISBN: 9780241556856.