# Meta-nudging honesty

*Past, present, and future of the research frontier*

Dimant, E.; Shalvi, S.

DOI
10.1016/j.copsyc.2022.101426
Publication date
2022
Document Version
Final published version
Published in
Current Opinion in Psychology

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Psychology

ELSEVIER

Review

# Meta-nudging honesty: Past, present, and future of the research frontier

Eugen Dimant[1,2] and Shaul Shalvi[3]

**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.

**Addresses**

[1] University of Pennsylvania, USA
[2] CESifo, Munich, Germany
[3] University of Amsterdam, the Netherlands

Corresponding author: Dimant, Eugen (edimant@sas.upenn.edu)

**Current Opinion in Psychology** 2022, **47**:101426

This review comes from a themed issue on **Honesty and Deception (2023)**

Edited by **Maurice E. Schweitzer** and **Emma Levine**

For complete overview about the section, refer Honesty and Deception (2023)

Available online 16 July 2022

https://doi.org/10.1016/j.copsyc.2022.101426



**Keywords**

Behavior change, Honesty, Lying, Nudging.

## 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

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

[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].

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 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 predisposition 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 article, 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.

## Conflict of interest statement

Nothing declared.

## References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest


1. Thaler RH, Sunstein CR: *Nudge: the final edition*. Penguin; 2021.
2. * DellaVigna S, Linos E: **RCTs to scale: comprehensive evidence from two nudge units**. *Econometrica* 2022, **90**:81–116.
3. Beshears J, Choi JJ, Laibson D, Madrian BC, Milkman KL: **The effect of providing peer information on retirement savings decisions**. *J Finance* 2015, **70**:1161–1201.
4. Bolton G, Dimant E, Schmidt U: **Observability and social image: on the robustness and fragility of reciprocity**. *J Econ Behav Organ* 2021, **191**:946–964.
5. Brandon A, Ferraro PJ, List JA, Metcalfe RD, Price MK, *et al.*: *Do the effects of social nudges persist? theory and evidence from 38 natural field experiments*. Working Paper; 2022.
6. * Gelfand M, Li R, Stamkou E, Pieper D, Denison E, *et al.*: **Persuading republicans and democrats to comply with mask wearing: an intervention tournament**. *J Exp Soc Psychol* 2022, **101**:104299.
7. Morvinski C, Saccardo S, Amir O: *Mis-nudging morality*. Manag Sci; 2022.
8. * Beshears J, Kosowsky H: **Nudging: progress to date and future directions**. *Organ Behav Hum Decis Process* 2020, **161**:3–19.
9. Dimant E: ***Hate trumps love: the impact of political polarization on social preferences***. Working Paper Available at SSRN: https://doi.org/10.2139/ssrn.3680871.
10. Hertwig R, Mazar N: *Toward a taxonomy and review of honesty interventions*. Working Paper; 2022.
11. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, *et al.*: **Evaluating the repli- cability of social science experiments in nature and science between 2010 and 2015**. *Nat Human Behav* 2018, **2**:637–644.
12. DellaVigna S, Pope D, Vivalt E: **Predict science to improve science**. *Science* 2019, **366**:428–429.
13. ** Dimant E, Clemente EG, Pieper D, Dreber A, Gelfand MJ: *Politicizing mask- wearing: predicting the success of behavioral*

*interventions among republicans and democrats in the u.s.*. Scient Rep; 2022.

14. Milkman KL, Patel MS, Gandhi L, Graci HN, Gromet DM, *et al.*: **A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appoint- ment**. *Proc Natl Acad Sci USA* 2021, **118**.

15. Köbis NC, Verschuere B, Bereby-Meyer Y, Rand D, Shalvi S: **Intuitive honesty versus dishonesty: meta-analytic evidence**. *Perspect Psychol Sci* 2019, **14**:778–796.

16. Hallsworth M, Kirkman E: *Behavioral insights*. MIT Press; 2020.

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

18. Sunstein CR: **The distributional effects of nudges**. *Nat Human Behav* 2022, **6**:9–10.

19. ** Dimant E, Gesche T: ***Nudging enforcers: how norm perceptions and motives for lying shape sanctions***. Working Paper Available at SSRN: https://doi.org/10.2139/ssrn.3664995.

20. ** Bicchieri C, Dimant E: *Nudging with care: the risks and benefits of social information*. Public Choice; 2019:1–22.

21. Dimant E: **Contagion of pro-and anti-social behavior among peers and the role of social proximity**. *J Econ Psychol* 2019, **73**:66–88.

22. * Bicchieri C, Dimant E, Gächter S, Nosenzo D: **Social proximity and the erosion of norm compliance**. *Game Econ Behav* 2022, **132**:59–72.

23. Dimant E: ***Distributions matter: measuring the tightness and looseness of social norms***. Working Paper Available at SSRN: https://doi.org/10.2139/ssrn.4107802.

24. Yip JA, Schweitzer ME: **Norms for behavioral change (nbc) model: how injunc- tive norms and enforcement shift descriptive norms in science**. *Organ Behav Hum Decis Process* 2022, **168**.

25. Fehr E, Gachter S: **Cooperation and punishment in public goods experiments**. *Am Econ Rev* 2000, **90**:980–994.

26. Balafoutas L, Nikiforakis N: **Norm enforcement in the city: a natural field experiment**. *Eur Econ Rev* 2012, **56**:1773–1785.

27. Gelfand MJ, Raver JL, Nishii L, Leslie LM, Lun J, *et al.*: **Differences between tight and loose cultures: a 33-nation study**. *Science* 2011, **332**:1100–1104.

28. * de Ridder D, Kroese F, van Gestel L: **Nudgeability: mapping conditions of suscep- tibility to nudge influence**. *Perspect Psychol Sci* 2021:1 745–691 621 995 183.

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

30. Andreoni J, Gee LK: **Gun for hire: delegated enforcement and peer punishment in public goods provision**. *J Publ Econ* 2012, **96**:1036–1046.

31. Bicchieri C, Dimant E, Xiao E: **Deviant or wrong? the effects of norm information on the efficacy of punishment**. *J Econ Behav Organ* 2021, **188**:209–235.

32. Bicchieri C, Dimant E, Sonderegger S: ***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://doi.org/10.2139/ssrn.3326146.

33. Deutscher C, Dimant E, Humphreys BR: ***Match fixing and sports betting in foot- ball: empirical evidence from the German Bundesliga***. Working Paper Available at SSRN: https://doi.org/10.2139/ssrn.2910662.

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

35. Cohn A, Fehr E, Marechal MA: **Business culture and dishonesty in the banking industry**. *Nature* 2014, **516**:86–89.

36. * Dimant E, Van Kleef GA, Shalvi S: **Requiem for a nudge: framing effects in nudging honesty**. *J Econ Behav Organ* 2020, **172**:247–266.

37. * Shalvi S, Dana J, Handgraaf MJ, De Dreu CK: **Justified ethicality: observing desired counterfactuals modifies ethical perceptions and behavior**. *Organ Behav Hum Decis Process* 2011, **115**:181–190.

38. Shalvi S, Gino F, Barkan R, Ayal S: **Self-serving justifications: doing wrong and feeling moral**. *Curr Dir Psychol Sci* 2015, **24**:125–130.

39. Saccardo S, Serra-Garcia M: *Cognitive flexibility or moral commitment? evidence of anticipated belief distortion*. Working Paper; 2022.

40. Golman R, Hagmann D, Loewenstein G: **Information avoidance**. *J Econ Lit* 2017, **55**:96–135.

41. Dimant E, Galeotti F, Villeval MC: *Norm-formation and the role of information acquisition*. 2022. Mimeo.

42. Vu L, Soraperra I, Leib M, van der Weele J, Shalvi S: *Willful ignorance: a meta- analysis*. Working Paper; 2022.

43. Mazar N, Ariely D: **Dishonesty in everyday life and its policy implications**. *J Publ Pol Market* 2006, **25**:117–126.

44. Weisel O, Shalvi S: **The collaborative roots of corruption**. *Proc Natl Acad Sci USA* 2015, **112**:10 651–710 656.

45. Leib M, Köbis N, Soraperra I, Weisel O, Shalvi S: **Collaborative dishonesty: a meta- analytic review**. *Psychol Bull* 2022, **147**:1241.

46. ** Weisel O, Shalvi S: **Moral currencies: explaining corrupt collaboration**. *Curr Opin Psychol* 2022, **44**:270–274.

47. Ren ZB, Dimant E, Schweitzer ME: *Social motives for sharing conspiracy theories*. Working Paper; 2022.

48. Ferrali R: *Is honesty or dishonesty more contagious? Evidence from the field*. Working Paper; 2020.

49. Zlatev JJ, Daniels DP, Kim H, Neale MA: **Default neglect in attempts at social influence**. *Proc Natl Acad Sci USA* 2017, **114**:13 643–13 648.

50. List J: *The voltage effect*. Penguin Books Limited; 2022, ISBN 9780241556856.