Title: Not all Interventions are Made Equal: Harnessing Design and Messaging to Nudge Bystander Intervention

Anna Davidovic[1], Adam Joinson[2], Catherine Hamilton-Giachritsis[3], and Othman Esoul[4]

[1] School of Management, University of Bath

[2] School of Management, University of Bath

[3] Department of Psychology, University of Bath

[4] School of Management, University of Bath

This is the accepted version of the following article: Davidovic A, Joinson A, Hamilton-Giachritsis C, et al. Not All Interventions and Made Equal: Harnessing Design and Messaging to Nudge Bystander Intervention. Cyberpsychology, Behavior, and Social Networking 2024; 0(0), doi:10.1089/cyber.2024.0223 which has now been formally published in final form at Cyberpsychology, Behaviour and Social Neworking at [https://doi.org/10.1089/cyber.2024.022]. This original submission version of the article may be used for non-commercial purposes in accordance with the Mary Ann Liebert, Inc., publishers' self-archiving terms and conditions.

Keywords: Bystander intervention; simulation; design; social media

Acknowledgements

We thank Joe Murphy for his analysis of the qualitative data for this project and we thank our funder for enabling this study to take place.

Author confirmation

Anna Davidovic: Conceptualization, Methodology, Resources, Data Curation, Formal analysis, Writing – Original Draft. **Adam Joinson**: Supervision, Writing – Reviewing and Editing, Funding acquisition. ***Catherine Hamilton-Giachritsis****:* Supervision, Writing – Reviewing and Editing. **Othman Osoul:** Software, Investigation, Data curation

Author disclosure

The authors have no conflicts of interest to disclose.

Funding statement

This study was part funded by EPSRC's National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN), grant number

Abstract

This study examined the influence of design 'nudges' on bystanders' willingness to intervene in online harassment using a social media simulation. Utilizing a 2X2 experimental design, we tested the ability of key design features (community guidelines and pop-up messaging) to induce a sense of self-efficacy (low/high) and personal responsibility (low/high) and thence to influence intervention levels. Participants (N=206) were invited to 'beta test' a new Social Networking Site (SNS) for 15 minutes. All participants were exposed to four instances of online harassment against a victim. Bystanders in the low efficacy and high responsibility condition were most likely to intervene, although this finding only applied to 'private' (e.g. direct, 1-2-1 messaging) rather than 'public' (e.g. posting on a public feed) interventions. Overall, participants preferred 'private' interventions which avoided public confrontation. Qualitative insights highlight a perceived lack of transparency in reporting options and a belief that interventions rarely made a difference as the 'damage had been done'. Results are discussed in relation to the amplification of personal responsibility when the SNS does not provide clear guidelines and reminders. We recommend ways of 'designing in' nudges in practice, to facilitate bystander intervention.

Keywords: Bystander intervention; simulation; design; social media; online harassment; self-efficacy

1. INTRODUCTION

Online harassment is a heavily debated term [1] consisting of a range of behaviours varying in intent, severity and harm. [2] Here, we define it as targeted abuse or harmful behaviour directed at another individual through computer-mediated communication (CMC). Exploring the role of the digital bystander is crucial since cyber-bullying (a form of online harassment) has been linked with poorer mental health outcomes, increased suicide ideation in young adults [3] and has been found to be more detrimental than traditional in-person bullying. [4]

Studies have identified numerous situational and personal factors that can facilitate intervention, such as prior relationship with the victim [5], anonymity [6] and level of personal responsibility. [7, 8, 9] Bystanders are more likely to intervene (both online and offline) when they are confident in their own abilities (a concept know as 'self-efficacy')[10] [5, 11] In contrast, those lower in self-efficacy act more passively in bystanding scenarios.[12] In the context of CMC, a study of young adults (n=1180) found that low self-efficacy reduced the likelihood of intervening in online hate speech.[8] There is promising support that self-efficacy can be increased through digital intervention [13], however there is large variance in the measurement of self-efficacy ranging from a stable, global sense of 'efficacy' or a more nuanced, situation specific efficacy with multiple facets. In this study, we examine both overall self-efficacy and bystander specific efficacy in encouraging intervention when witnessing cyber bullying.

In recent years, there has been an innovative shift towards to use of simulation paradigms to measure bystander behaviour [14, 15, 16, 17] and 'designing in' bystander intervention through "nudges". [18] Digital nudges are interventions that steer people while also allowing them to "go their own way" (p.3). [19] Specifically, we test whether 'reinforcement nudges' (increasing the salience of behaviour in the mind of the user) and social nudges (inducing a social norm) [19] can increase bystander intervention. This study seeks to leverage the concepts of self-efficacy and personal responsibility specifically since they are well

supported by existing literature. We take existing research [15] one step further by expanding the range of intervention options available to the bystander (as recommended by recent qualitative research [20]) in the both the experimental design and analysis. The research has three interrelated research questions:

RQ1: To what extent do self-efficacy and personal responsibility 'nudges' increase bystanders' likelihood of intervening in online harassment?

RQ2: What is the interaction between personal responsibility and self-efficacy in relation to bystander intervention?

RQ3: How do efficacy and responsibility nudges influence the t*ype* of intervention (e.g. private vs. public) used to tackle online harassment?

2. MATERIALS AND METHODS

A bystander simulation was conducted in 2023 using an adapted version of 'SnapEatLove'; a high validity open-source platform [15] which mimics the capabilities of the SNS Instagram (e.g. sharing, liking, posting) with a focus on food (see Supplementary material 2). The platform was developed for studying behaviour in social network sites in a naturalistic manner. We adapted the simulation to include a wider range of bystander intervention options [20] and varied the location of 'other users' to cities in the United Kingdom. All other user-interactions remained the same as per the original simulation (see procedure).

2.1 Design

The study adopted a 2 (low efficacy / high efficacy) X 2 (low responsibility / high responsibility) between-subjects factorial design. Self-efficacy and sense of personal responsibility were manipulated through two key design features: 1) the "community guidelines" prior to log on, and 2) 'pop up' messaging on their newsfeed (see Supplementary 3). Participants were required to manually 'close' the message which ensured they attended to the information. To induce a sense of responsibility we used a 'social norm nudge' stating

it was the responsibility of individual users to keep the community safe (high) or the responsibility of moderators (low). Self-efficacy was induced through 'reinforcement nudges' whereby users were reminded of all intervention options available for tackling online harassment (high) or omitting this information (low).

2.2 Participants

The participants (*n*=206, 49.5% Female, 49% Male and 0.5% Non-binary/third gender) were recruited using participant platform Prolific, under the guise of beta testing a new social media platform. Prolific screening was set that all participants were based in the UK and ensured a balanced gender composition. Under 16's were excluded from the sample. Participants were removed from the study if any of the following criteria were met: 1) they did not create a profile on the site, 2) they did not create one post, 3) they did not interact with others at least once (e.g. comment, like/share, report). Participants were paid above the national living wage at the time of the experiment (£9.50 per hour).

2.3 Ethics

Full ethical approval was granted through the [name removed] ethical committee. Given the deception involved in the study, mitigations were taken; 1) participants were fully debriefed on the aims following their participation, 2) they were provided with the opportunity to anonymously withdraw their data within one week, and 3) were signposted to cyberbullying resources.

2.4. Procedure

Participants were asked to provide information about their current social media use, demographics (e.g. age, gender) and directed to the SNS ('SnapEatLove') where they created profile information for realism purposes. Over 15-minutes, participants were exposed to a feed of interactions occurring in apparently real time alongside other users (who were in

reality pre-programmed bots) where they could share / flag / post as they would normally in their day to day lives. All participants were exposed to four instances of online harassment:

INSERT TABLE 1

2.5 Tracking Intervention Rates and type

All interactions were logged and captured in individual CSV files (except images for ethical reasons). Overall bystander intervention was the sum of all 'interventions' directed at the four harassment messages (our experimental stimuli) only (see Table 1).

2.6. Measures

After their allocated time on SnapEatLove, participants were directed to a survey site to complete a follow up questionnaire before debriefing (full detail in Supplementary material 1).

2.7. Experimental stimuli

Participants were exposed to four instances of online harassment against a single victim by a single offender. We used the same four posts as used by a previous study [15] which had been tested for realism and severity (Table 1).

2.8. Analytic strategy

Multiple analysis of variance (MANOVAs) were used since they would allow us to test the impact of all both main effects (responsibility and efficacy) as well as interactions, and multiple intervention outcomes (and co-variates) simultaneously. Similarly, this negates the need to adjustments for multiple analyses to account for the slight underpower of our sample size (n=206). The recommended sample size was calculated as 267 participants using GPower Software for a medium effect size according to Cohen (1998) and this should be factored into the interpretation of our results. We used the widely adopted p-value of 0.05 throughout to determine the impact of all main effects and interactions.

3 RESULTS

3.1 Manipulation check

We first conducted a manipulation check of the experimental conditions using a MANOVA. We included overall general self-efficacy score (M = 31.19, SD = 4.56) and a one-item responsibility score into the model as covariates.

Firstly, we examined the self-efficacy manipulation on intervention specific process and outcome efficacy. Level of confidence in one's ability to intervene ('process efficacy' measured on a 5-point Likert scale) was significantly different between the low and high self-efficacy conditions, $F(1,182) = 6.06$, $p = 0.015$. Specifically, those in the 'low efficacy' scored significantly lower on confidence (M = 4.19, SD = 0.86) than the 'high efficacy' condition (M = 4.43, SD = 0.71) suggesting this manipulation was effective. There was no significant difference between groups in relation to confidence that this intervention would make a difference (outcome efficacy, $F(1,182) = 0.00$, $p = 0.96$). Regarding the responsibility manipulation, self-reported responsibility was not significantly different between groups, suggesting this intervention alone was not effective.

However, when examining the interaction between both conditions, we found that there was a significant difference between groups in self-reported levels of responsibility, $F(1,182) = 5.92$, $p = 0.016$. Specifically, users' reported sense of responsibility on the platform was highest in the condition combining low efficacy / high responsibility (M = 4.07, SE = 0.15) compared to low efficacy / low responsibility (M=3.50, SE =1.44). Therefore, for the remainder of the analysis, we will focus on exploring interaction effects rather than the main effects of the conditions.

3.2 The role of responsibility and efficacy 'nudges' on overall intervention rates

We explored whether the presence of personal responsibility 'nudges and efficacy 'nudges' increased bystander intervention on SnapEatLove. A between-measures MANOVA (using the same covariates) revealed that induced efficacy alone was not a significant

predictor, $F(2, 206) = 0.07$, $p = 0.79$, nor was induced responsibility, $F(2, 206) = 2.64$, $p = 0.11$. However, the interaction between both nudges on intervention rates was approaching significance, $F(2,206) = 3.58$, $p = 0.06$. Specifically, participants intervened most in the low efficacy / high responsibility condition (M = 2.24, SE = 0.21) in comparison to the low efficacy / low responsibility condition (M = 1.44, SE = 0.21).

3.3 Other drivers of overall intervention

A linear regression examined intervention scores based on age. Age significantly predicted intervention scores, $\beta = -.18$, $p = 0.01$ with intervention reducing as age increased. Gender did not have a significant impact on intervention, $t(201) = 0.31$, $p = 0.6$.

3.4 The impact nudges on type of intervention – public vs. private

Rates of intervention were higher than expected, with only 28% of participants choosing not to intervene in any way (non-intervention rates were 75.4% in a previous study [15], although a more limited range of intervention options were measured). Interestingly, the preferred intervention option was 'liking' the victims original post (31.1 % liked one post, 10.7% liked two posts) followed by flagging bully comments (21.4% flagged one post, 13.1% flagged twice).

To explore how design nudges impacted choice of intervention, we created two new variables: 1) number of 'public' interventions; and 2) number of 'private' interventions, for each participant. In line with previous research, [11] private interventions (M = 1.016, SE = 0.41) were preferred over public interventions (M = 0.800, SE = 0.29). A between-measures MANOVA determined the impact of experimental condition on type of intervention. Taking each main effect in turn, there was no significant difference between the likelihood of conducting a public or private intervention. However, taken together as an interaction, the nudges significantly increased private interventions, $F(1,182) = 7.76$, $p = 0.01$ but interestingly had no effect on public interventions, $F(1,182) = 0.22$, $p = 0.64$. Specifically,

those in the low efficacy / high responsibility condition exhibited the highest rates of private intervention *(M* = 1.46, *SD* =0.16) compared to low efficacy / low responsibility (*M* = 0.71, *SD*=0.16).

3.5 Qualitative insights

We examined all qualitative, free-text data collected in the follow-up survey through a process of bottom up, semantic coding following the practices set out in Thematic Analysis.[21] (See supplementary material 4 for full analysis). A strong theme was that social media companies themselves (or moderators) are responsible rather than individual users. In relation to self-efficacy, participants noted a lack of transparency and a perception that their intervention would not be effective. Exploring this further, participants described that the damage (social, emotional, reputational) to the victim had already been done.

4 DISCUSSION

The present research explored how efficacy and responsibility design 'nudges' influence bystander intervention in a realistic simulation. We found that self-efficacy 'nudges' were effective and could be easily operationalised in the future, unlike responsibility 'nudges'. This could speak to the difficulty of inducing a sense of responsibility when the bystander is immersed in a virtual network of acquaintances rather than close friends (a consistent finding in the broader literature [5]).

Taken on their own, efficacy and responsibility 'nudges' did not increase overall bystander intervention. More promising results were found when nudges were combined, although not statistically significant. In future studies, larger sample size and thus greater power may validate these findings. When examining the *types* of intervention, we found that a combination of the two design nudges (low efficacy / high responsibility) increased the likelihood of 'privately' intervening (e.g. DM's, reporting) but not public interventions (e.g. public reply). This finding suggests that a sense of responsibility *can* drive intervention under

the right circumstances. One possible theoretical explanation is that the ambiguity created by *not* giving information on how to intervene (in the low efficacy condition) further amplified feelings of personal responsibility (in the high responsibility condition). In other words, if bystanders are not confident in intervention options offered by the SNS – they take intervention into their own hands (e.g. DM's). In line with previous studies, participants preferred 'private' interventions such as comforting victims, that do not typically rely on the platform itself and often lack confidence in the SNS reporting process.[20]

In this study, we aimed to increase self-efficacy by clearly articulating and reminding participants of the intervention options. This manipulation alone was not sufficient to increase intervention (although it was a key covariate in our interaction analysis). Our qualitative findings suggest that bystanders often felt that interventions are unlikely to make a difference as the 'damage had been done'. As such, we recommend future research and interventions could focus on the *benefits of intervention* (e.g. where intervention can lead to positive change), targeting the motivation to act in the first place. Our qualitative findings also suggest that bystander confidence (and the perceived social ramifications of 'making it worse') was a key barrier, regardless of intervention condition. Thus, we recommend future research and campaigns could specifically target 'social' self-efficacy. Previous research has highlighted the specific role of 'social' self-efficacy; [22] one's confidence in social interaction and their perceived ability to resolve conflicts.

5. CONCLUSIONS

This study sheds light on the combined role of efficacy and responsibility on bystander intervention and supports the notion that such interventions can be 'designed into' SNS using a realistic and controlled simulation. A key limitation is that participants were interacting with unknown users (bots) not their own social network community, which may well limit their overall sense of responsibility. Nevertheless, the study provides insight into

how design initiatives can impact different types of digital intervention (e.g. public vs. private). Specifically, this study supports the idea that SNS should ensure bystanders have ample opportunities to privately intervene [8]. Our findings also suggest that interventions should target bystander confidence rather than bystander ability, for example through mandatory bystander education.

6. ACKNOWLEDGMENTS

We thank JM for their analysis work on the qualitative data for this project and our funders.

7. AUTHOR CONTRIBUTIONS

AJ led on the funding acquisition for the project. AD devised the study (conceptualisation, methodology, theoretical framework) with input and supervision from AJ and CH. OA led on the data curation and software for the simulation with input from AJ and AD. OA and AJ conducted the simulation and collected the data. AD cleaned and analysed the data and wrote the manuscript with support from AJ and CHG. AJ and CH reviewed the manuscript.

8. AUTHOR DISCLOSURE

The authors declare there are no conflicts of interest.

9. FUNDING STATEMENT

This research was part funded by the EPSRC's National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN), grant number: EP/V011189/1.

10. TABLES

Table 1

Online Harassment Posts Used in the Simulation (previously used and tested by Difranzo (2018)

Post 1	*"Stop posting this shit, nobody cares."*
Post 2	*"Your life is sad. Look at what you eat."*
Post 3	*"When will you get it into your head that nobody likes your stupid ass."*

Post 4	*"This photo is uglier than you, and that's saying something."*

11. FIGURES

N/A

12. SUPPLEMENTARY FILES

These are uploaded as an individual file

REFERENCES

1. Marwick AE. Morally Motivated Networked Harassment as Normative Reinforcement. Social media + society 2021;7(2):205630512110213, doi:10.1177/20563051211021378
2. PewResearchCentre. Online Harassment Survey. 2021. Available from: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.
3. John A, Glendenning AC, Marchant A, et al. Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. Journal of Medical Internet Research 2018;20(4), doi:10.2196/jmir.9044
4. Perren S, Dooley J, Shaw T, et al. Bullying in school and cyberspace: Associations with depressive symptoms in Swiss and Australian adolescents. Child and Adolescent Psychiatry and Mental Health 2010;4(doi:10.1186/1753-2000-4-28
5. Bastiaensens S, Vandebosch H, Poels K, et al. Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. Comput Hum Behav 2014;31(1):259-271, doi:10.1016/j.chb.2013.10.036
6. Andalibi N, Forte A. Responding to sensitive disclosures on social media: A decision-making framework. ACM Transactions on Computer-Human Interaction 2018;25(6), doi:10.1145/3241044
7. Latané BD, J. M. The Unresponsive Bystander. Appleton-Century-Croft: New York, NY; 1970.
8. Obermaier M. Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. New Media and Society 2022, doi:10.1177/14614448221125417
9. Leonhard L, Rueß C. Obermaier, M., & Reinemann, C. Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. Studies in Communication and Media 2018;7(4):555-579, doi:10.5771/2192-4007-2018-4-555
10. Bandura A. Self-efficacy: Toward a unifying theory of behavioral change. Psicoterapia Cognitiva e Comportamentale 2022;28(1):81-104
11. DeSmet A, Bastiaensens S, Van Cleemput K, et al. Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. Comput Hum Behav 2016;57(398-415, doi:10.1016/j.chb.2015.12.051
12. Thornberg R, Wänström L, Hong JS, et al. Classroom relationship qualities and social-cognitive correlates of defending and passive bystanding in school bullying in Sweden:

A multilevel analysis. Journal of School Psychology 2017;63(49-62, doi:10.1016/j.jsp.2017.03.002
13. Kutok ER, Dunsiger S, Patena JV, et al. A cyberbullying media-based prevention intervention for adolescents on instagram: Pilot randomized controlled trial. JMIR Mental Health 2021;8(9), doi:10.2196/26029
14. Taylor SH, Difranzo D, Choi YH, et al. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. Proceedings of the ACM on Human-Computer Interaction 2019;3(CSCW), doi:10.1145/3359220
15. Difranzo D, Taylor SH, Kazerooni F, et al. Upstanding by design: Bystander intervention in cyberbullying. 2018.
16. Markey PM. Bystander intervention in computer-mediated communication. Comput Hum Behav 2000;16(2):183-188, doi:10.1016/S0747-5632(99)00056-4
17. Dillon KP, Bushman BJ. Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. Comput Hum Behav 2015;45(144-150, doi:10.1016/j.chb.2014.12.009
18. Masur PK, DiFranzo D, Bazarova NN. Behavioral contagion on social media: Effects of social norms, design interventions, and critical media literacy on self-disclosure. PLoS ONE 2021;16(7 July), doi:10.1371/journal.pone.0254670
19. Bergram K, Djokovic M, Bezençon V, et al. The Digital Landscape of Nudging: A Systematic Literature Review of Empirical Research on Digital Nudges. 2022.
20. Davidovic A, Talbot C, Hamilton-Giachritsis C, et al. To intervene or not to intervene: young adults' views on when and how to intervene in online harassment. Journal of Computer-Mediated Communication 2023;28(5), doi:10.1093/jcmc/zmad027
21. Braun V, & Clarke, V. Succesful Qualitative Research: A Practical Guide for Beginners. SAGE Publications: London; 2013.
22. Yang L, Gao T. Defending or not? The role of peer status, social self-efficacy, and moral disengagement on Chinese adolescents' bystander behaviors in bullying situations. Current Psychology 2023;42(33):29616-29627, doi:10.1007/s12144-022-04039-1