Complacent Agents: How Transparent Nudging Functions in a Controversial Context

Abstract:

Transparent nudges have been shown to be an effective behavior change intervention, especially for choice architects with prosocial goals. Yet, little empirical research has investigated how they function when used by choice architects who are not perceived as obviously prosocial. In this three-group between subject study (N = 310), I examine donation rates and perceived freedom scores across both a transparent, and a traditional default nudge for a controversial charity. The transparent nudge explicitly informs users of the presence and effect of the default, while the traditional nudge intervenes without warning users. Both nudge types prove to be equally effective at inducing high donations. Contrastingly, those in the transparent nudge condition reported lower perceived freedom scores than those in the traditional nudge condition. Participants who felt neutral about the goals of the charity donated lowest overall, and had the highest perceived freedom scores. These results suggest that both nudge type, and agreement may have a stronger influence on efficacy and perceived freedom than what was previously expected.

Keywords: transparent nudging, behavior change intervention, libertarian paternalism

By Jordan Selesnick

Preregistration and OSF files

I. Introduction

The ethical justifications of behavior change interventions must constantly be reexamined when applied to new contexts. Perhaps the most popular intervention type, commonly referred to as 'nudging,' has been central to understanding the ethics of behavior change interventions. This is because nudges are derived from the seemingly oxymoronic idea of libertarian paternalism; they influence users by altering the structure of their choices while simultaneously preserving their freedom to choose otherwise. (Thaler & Sunstein, 2003).

Paternalism and libertarianism effectively exist on a spectrum. Paternalists are outcome oriented, where deliberate structure is used to create behavior change. Libertarians are rule oriented, in a deontological sense. For an intervention to be libertarian, it must preserve users freedom to make whatever choice they desire. While it is possible to achieve both goals in a behavior change intervention, they are typically in conflict with each other. For example, a nudge that uses a default makes the tradeoff of slightly limiting freedom (because the non-default choices require more effort to choose), in order to increase the likelihood of achieving their goal (Hausman & Welch, 2010).

It is easy to understand the justifications for promoting libertarian leaning interventions. In line with Johnathan Haidt's moral modules for intuitive ethics, libertarianism is celebrated because it maximizes fairness in a way that is obvious for observers to understand (Haidt & Joseph, 2004). Paternalism on the other hand, has been traditionally dubious given its potential for exploitment. Those who defend paternalistic leaning interventions, often cite that the cost of limiting freedom is ethically justified because nudging people to make good choices prevents future harm (Guala & Mittone, 2015). Similarly, they state that because choice architecture is omnipresent (even if it is not always deliberate), it is preferable to have a benevolent choice architect intervene, rather than let people be influenced by random circumstances. For this claim to be true, it requires that the choice architect's belief about what is best for the user is accurate, an issue at the heart of the subsequent study. While this framework only outlines the surface of the ethics in nudging debate, it shows that institutions have the ability to turn the libertarian-paternalism dial such that their behavior change interventions match the goals of their situation (Schmidt & Engelen, 2020).

In the subsequent study, I aimed to examine efficacy and perceived autonomy across nudge type (traditional vs. transparent), and agreement type (agree vs. disagree vs. neutral) using a between subject design. In line with theories outlined in the literature, I expected implicit recommendation and psychological reactance to dictate efficacy based on agreement type. I also expected that perceived autonomy would be mediated by agreement, especially when participants were aware of the nudge.

II. Literature Review

In an effort to explore ways in which choice architects could reduce paternalism in behavior change interventions, there has been a growing body of research on transparent nudging – the practice of explicitly telling users that they are being nudged (Bruns et al., 2018). Choice architects typically do this by including transparency statements that indicate that their nudge is powerful (it influences decisions), and purposeful (it helps achieve their goals).

Transparent nudges change the dynamics of nudging in two crucial ways. First, they effectively make salient the implicit conversation that happens between choice architect and user (Krijnen, Tannenbaum, & Fox, 2017). Nudges act as recommendations; a default is a signal to users that the choice architect believes the default is a good choice for the user to make. Transparent nudges make users think about this typically subconscious process in an active way. Second, they alert users to the fact that they are being influenced, which has the potential to cause psychological reactance (Bruns & Perino, 2019).

In spite of previous concerns, transparent nudges have proven to be successful from an efficacy standpoint. Historically, nudges have been criticized to 'only work in the dark,' meaning that transparency would have a negative impact on their effectiveness (Ivanković & Engelen, 2019). A number of studies have shown that this is not true; transparent nudges led to similar outcomes when compared to traditional, opaque alternatives (Bruns et al., 2018).

Efficacy does not capture the entire picture alone. The impetus for transparent nudging was to meet ethical concerns and improve user wellbeing, therefore it is imperative to examine outcomes from the user experience perspective. Previous research has suggested that people

perceive defaults as unethical due to their potential for limiting user autonomy (Hagman, Andersson, Västfjäll, & Tinghög, 2015). Yet, until recently, these claims have had little empirical support. Patrik Michaelsen has spearheaded an effort to change this through a series of papers that analyzed perceived autonomy, and choice satisfaction among participants who experienced a variety of nudges (Michaelsen, Johansson, & Hedesström, 2021). Participants scored highly in studies using between subject designs, and lower on ones using within subject designs. This was likely because transparency made them aware that they were potentially being manipulated (Michaelsen et al., 2020). This result is valuable because it sheds light on a tradeoff of transparency; in an effort to be more ethical and provide users with informational freedom, those same users perceive themselves as less free.

Michaelsen's findings open up a discussion about other contextual factors that could influence efficacy and perceived autonomy in transparent nudging. A notably understudied issue in the current state of the literature is how institutional intention aligns with user intention. The majority of the literature is focused on how transparent nudging functions in generally prosocial contexts (Schmidt & Engelen, 2020). Yet, in practice nudges are often used in a variety of areas where the ethics of the choice architect may not be as one dimensional. As noted previously, paternalism was only considered justified when the choice architect's intentions matched the goals of the user. Under conditions of a traditional nudge, users may not have been able to identify situations where goals did not align. Transparent nudges change the dynamic here – they ensure that users recognize what is happening so that they can judge alignment themselves.

III. Experimental Design

For this study, I used a three-group between-subject design with the aim of exploring the relative effectiveness of a transparent nudge for a controversial charity, as compared to a traditional nudge and a control. This experiment was conducted via an online Qualtrics survey with participants recruited from Amazon Mechanical Turk (MTurk).

A crucial element for success was choosing a charity that matched my goals. It was necessary to use a charity that was controversial enough such that people would reasonably agree, remain neutral about, or disagree with their mission. The charity also could not be too controversial,

where participants' beliefs would be too strong to be nudged (Sunstein, 2017). After a series of pretests, I eventually chose The Salvation Army. They fit my needs because they are well known, have generically altruistic goals across a variety of contexts (feeding poor, disaster relief, etc.), and have a somewhat dubious history of discrimination due to strong religious backing.

At the start of the study, participants were awarded 10 game tokens. For incentive compatibility, participants were told that each game token functioned as a raffle ticket for a $20 Amazon gift card, meaning the more tokens they kept, the more likely they were to win. After, participants were given a brief synopsis on the Salvation Army (see Appendix A) that remained neutral. It stated the charity's goals, accomplishments, and alleged history of discrimination. Participants were then given the opportunity to donate any number of their tokens to The Salvation Army. Next, participants were presented with a donation slider ranging from 0 to 10, set to a default donation of 5 tokens. Once the donation decision was complete, participants answered a series of questions that assessed their feelings of autonomy and perceived freedom, based on questions from previous studies (Michaelsen et al., 2020). Then, participants stated whether they agreed, felt neutral about, or disagreed with the goals of The Salvation Army. The survey concluded with demographic questions about gender, age, and race.

Treatments

Participants were randomly assigned to one of 3 conditions:

(a) Salvation Army Transparent
(b) Salvation Army Traditional
(c) Control

Salvation Army Transparent: Participants were presented with the opportunity to donate any number of their tokens to the Salvation Army. Above their donation slider was a transparency statement that read: "Please consider that the preselected default value may have an influence on your decision. This is meant to encourage higher donations to the Salvation Army" (Appendix D).

Salvation Army Opaque: Participants were presented with the opportunity to donate any number of their tokens to the Salvation Army. There was no transparency statement in this treatment.

Control: Participants were presented with the opportunity to donate any number of their tokens to an unnamed, unspecified charity. There was no default and no transparency statement in this group.

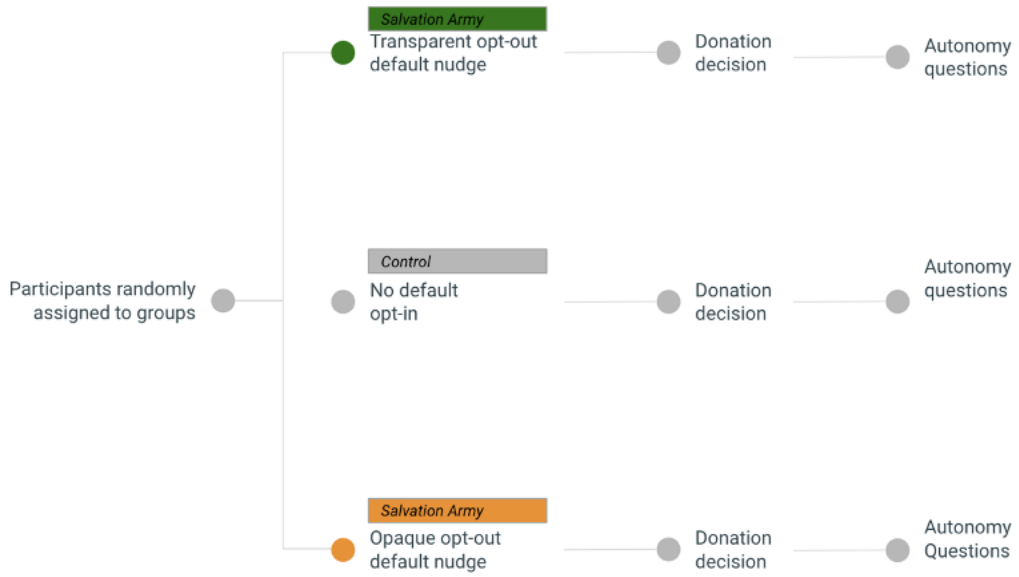


Figure 1.

IV. Hypotheses

My hypotheses were derived from two main points found in the literature. First, transparent nudges have been proven to have a similar effect as their traditional counterpartFs (Schmidt & Engelen, 2020). Second, it has been claimed that transparent nudges could theoretically lead to psychological reactance, indicating that those who disagree with the choice architect would likely have low contributions (Bruns & Perino, 2019). While Michaelsen et al. only found low perceived freedom scores for within subject design studies, I expect that the psychological reactance from those who disagree with the choice architect's goal will lead to similar scores in

this between subject study (Michaelsen et al., 2020). For those unaware of the nudge, I expect no infringement on perceived autonomy. As such, my hypotheses are as follows:

Contribution Hypotheses

H1: Both nudge conditions will have higher donation rates than the control.

H2: Donation rates will not differ between the transparent and traditional nudge conditions.

H3: Donation rates for those who disagree with, or are neutral about the choice architect will be lower than contributions rates for those who agree with the choice architect.

Perceived Autonomy Hypotheses

H4: Perceived autonomy scores for the transparent nudge condition will be lower than perceived autonomy scores for the control.

H5: Perceived autonomy scores for the transparent nudge condition will be lower than perceived autonomy scores for the traditional nudge condition.

H6: Perceived autonomy scores for those who disagree with, or are neutral about the choice architect will be lower than perceived autonomy scores for those who agree with the choice architect.

Sample

The experiment was pre-registered on aspredicted.org and was executed using a Qualtrics survey distributed through MTurk to N = 392 participants, all of whom were paid at a rate of $5.2 per hour for completing the survey. One participant was awarded a $20 gift card for winning the raffle. All sampled participants had completed at least 500 tasks on MTurk with a 99% completion rate. Due to failed attention checks, I excluded 82 participants, leaving a final sample of N = 310. According to a 2-tailed A Priori Wilcoxon-Mann-Whitney Test run on G*Power using pretest data, I would have needed 166 participants per treatment group to reach 80% power. While I acknowledge that running a well-powered study would have been ideal, due to budget constraints, I ended up with lower amounts of participants per treatment group. As such, I will not assume normalcy and use the median results of my dependent variables for all tests.

V. Empirical Analysis and Results

Dependent Variables

The first dependent variable was donation rate. The second dependent variable was freedom score, derived from a series of 8 total perceived autonomy and freedom questions, as established by Michaelsen (Michaelsen, Johansson, & Hedesström, 2021). These questions all use 11-point Likert scales (0-10) to assess participants' feelings (Appendix E). Perceived autonomy and freedom scores were averaged for each participant, which generated an overall freedom score.

Independent Variables

The first independent variable was treatment type. The second independent variable was 'agreement with the goals of the choice architect.' Participants were categorized as 'agree,' 'disagree,' or 'neutral' according to their survey responses.

Summary statistics

Summary statistics for conditions can be found in table 1 below. Summary statistics for agreement type (from within the nudge conditions) can be found in table 2 below.

	Control	*Transparent Condition*	*Traditional Condition*	*Overall*
Median Donation	2	7	7	5
Mean Donation	2.47	5.18	5.26	4.32
Median Perceived Freedom	5.63	5.25	5.56	5.38
Mean Perceived Freedom	6.42	6.03	6.67	6.37
N	101	107	102	310

Table 1.

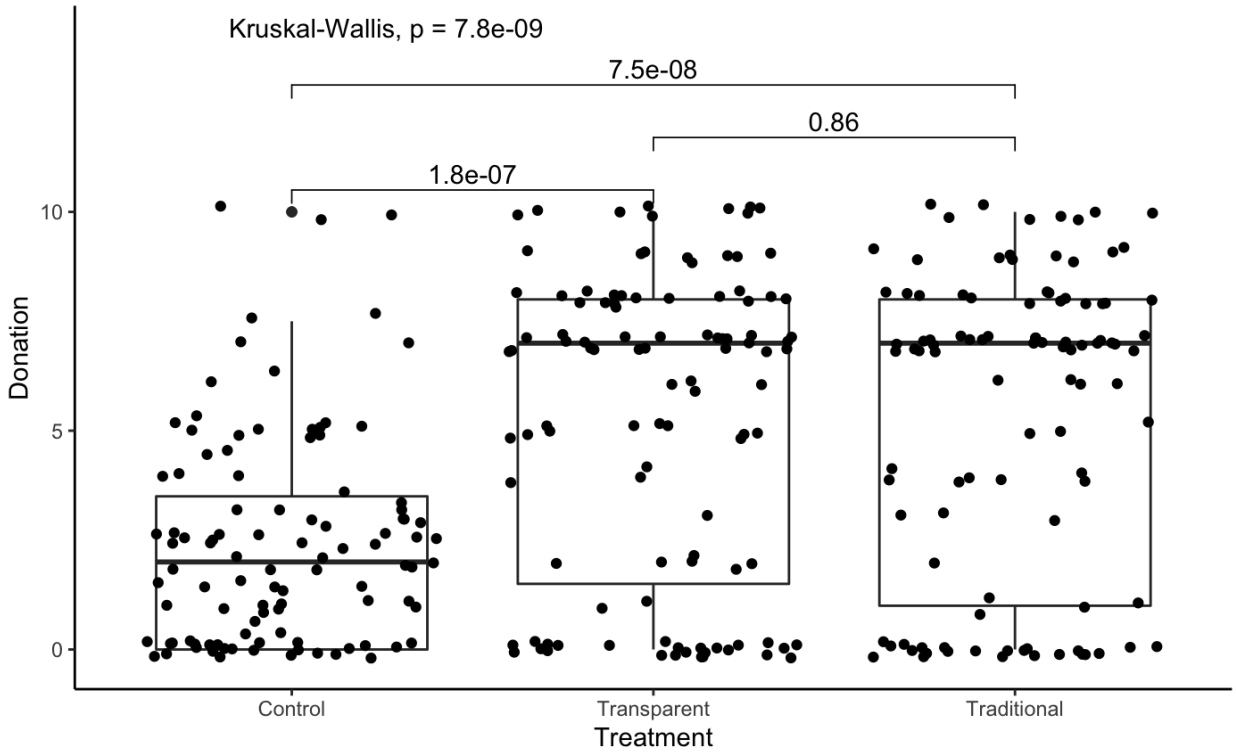
	Agree	*Disagree*	*Neutral*	*Overall*
Median Donation	7	7	0	7
Mean Donation	6.06	5.1	1.94	5.22
Median Perceived Freedom	5.25	5.13	8.63	5.38
Mean Perceived Freedom	6.09	5.85	7.74	6.34
N	144	29	36	209

Table 2.

Impact of Nudges on Donation Rates (H1 & H2)

The literature has shown that nudges (both traditional and transparent) are highly effective at inducing behavior change. A goal of this study was to determine whether nudges were impactful

for influencing donation rates to a controversial charity. In this study, participants in both treatment conditions donated significantly more than participants in the control group. Figure 2 shows that the median donation of both the traditional (M = 7) and the transparent (M = 7) nudge group was significantly higher than that of the control group (M = 2) with Cohen's Ds of 0.93 and 0.90, respectively. There was no significant difference in donation rates between the traditional and transparent nudge conditions. These results indicate that nudges are indeed effective, even when transparent. They also suggest that transparency does not have an effect on nudge efficacy in this context.



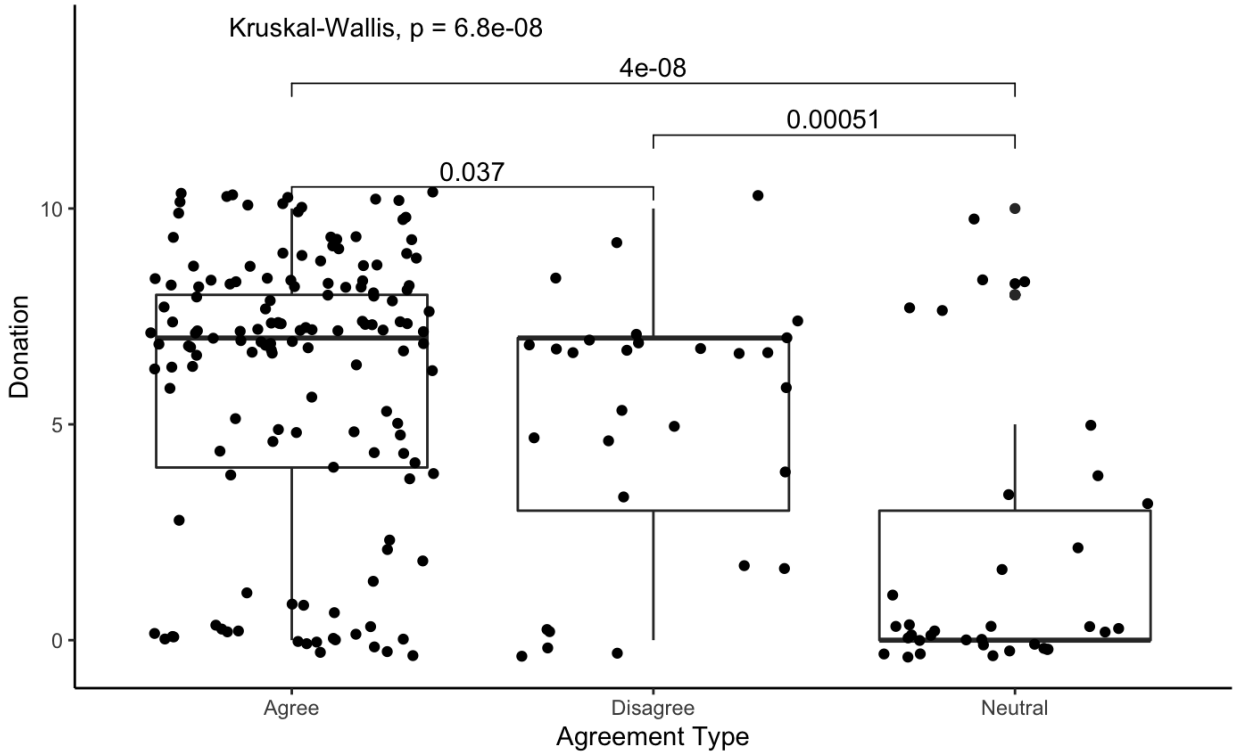
P-values are from Wilcoxon rank sum tests.

Figure 2.

Impact of Agreement on Donation Rates (H3)

Another goal of this experiment was to investigate if participant agreement impacted donation rates. I found significantly different results across all three comparisons. As shown in Figure 3, participants who agreed (M = 7) donated slightly more than those who disagreed (M = 7), with a small Cohen's D of -0.31. Participants who agreed (M = 7) donated much more than those who

were neutral (M = 0), with a large Cohen's D of -1.3. Finally, those who disagreed (M = 7) donated much more than those who were neutral (M = 0), with a large Cohen's D of -1.04. Interestingly, they show that those who felt neutral contributed much less than everyone else.

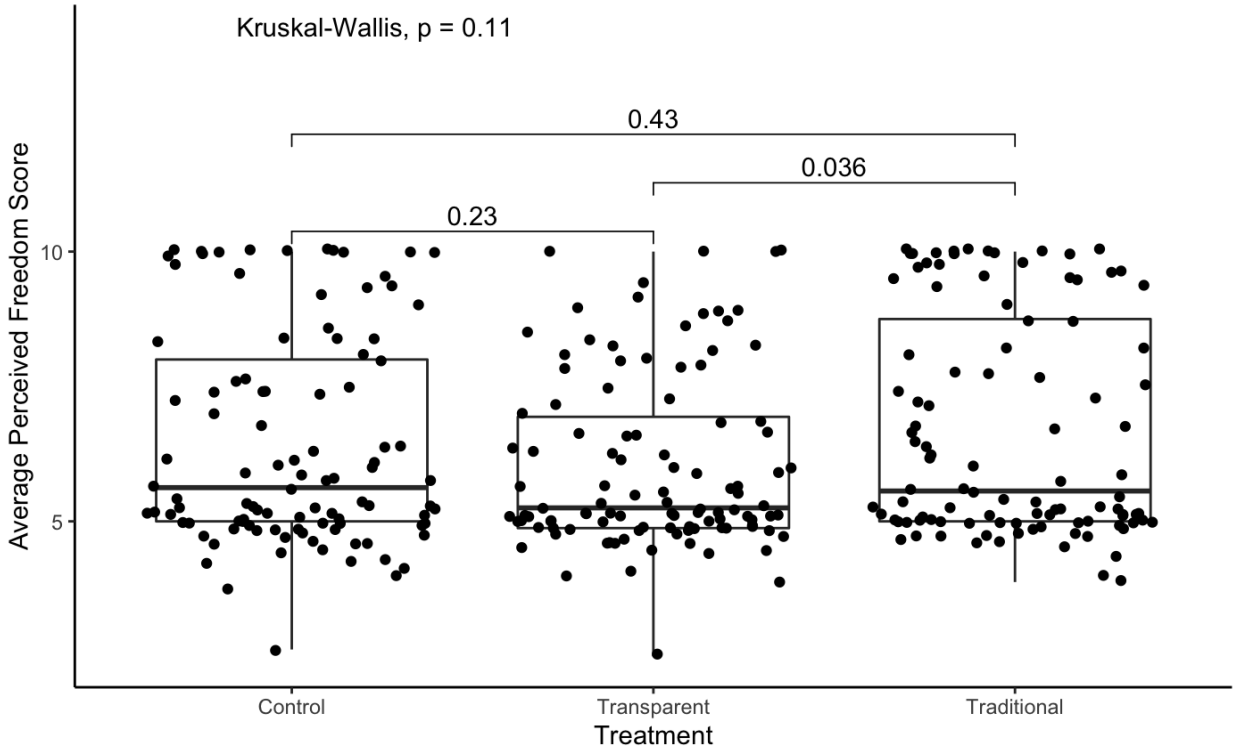


P values are from Wilcoxon rank sum tests.

Figure 3.

Impact of Nudges on Perceived Freedom Scores (H4 & H5)

The literature has shown that transparent nudges can have an adverse effect on perceived freedom scores. Figure 4 shows that those in the transparent nudge condition (M = 6.03) had lower perceived freedom scores than those in the traditional nudge condition (M = 6.67), with a Cohen's D of 0.35. I found no other significant difference between conditions. This result replicates that transparency may have an adverse impact on perceived freedom.

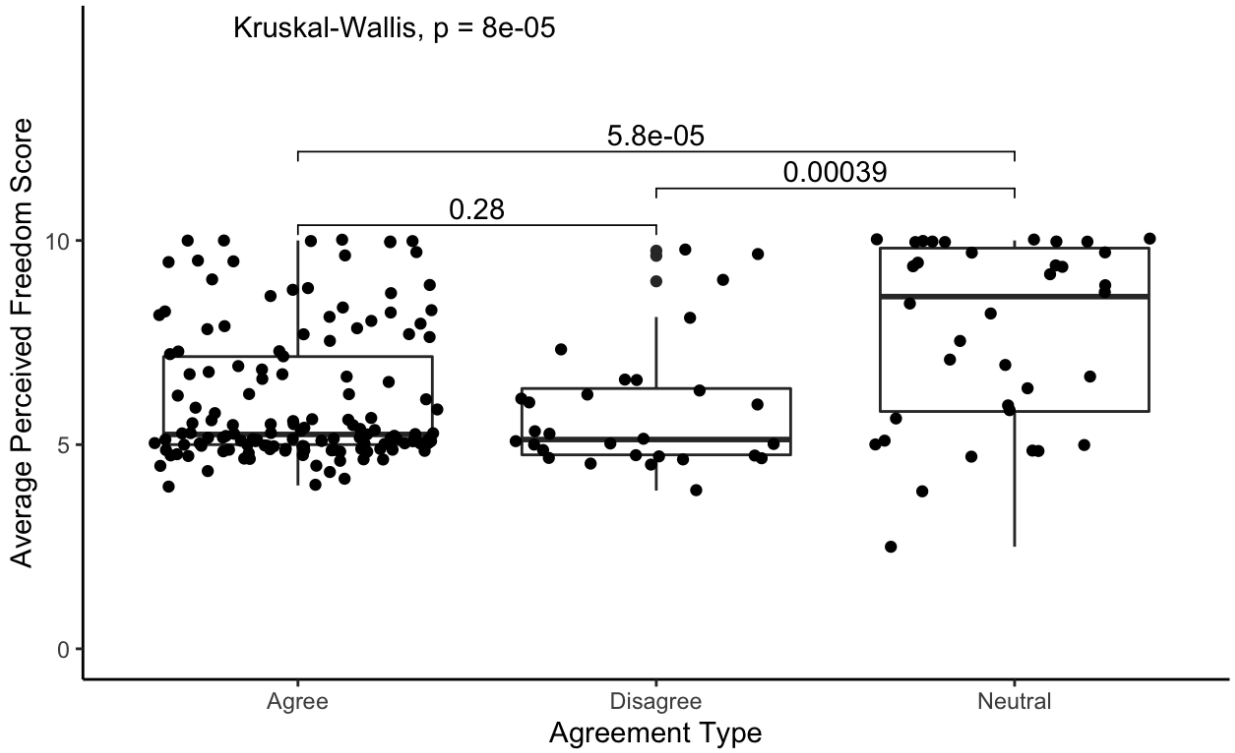


P-values are from Wilcoxon rank sum tests.

Figure 4

Impact of Agreement on Perceived Freedom Scores (H6)

A novel aspect of this study was analyzing how agreement affects perceived freedom across different types of nudges. Figure 5 shows that the median perceived freedom scores for both those who agree (M = 5.25) and those who disagree (M = 5.13) were significantly lower than those who were neutral (M = 8.63), with Cohen's Ds of 0.93 and 0.98, respectively. There was no significant difference in perceived freedom scores between those who agreed and those who disagreed. This result continues to show the unique behavior of participants who were neutral.



P values are from Wilcoxon rank sum tests.

Figure 5.

Sensitivity

To assess sensitivity, regressions using demographic variables were conducted on both donation rates and freedom scores. The analyses yielded very similar significance levels across both metrics, indicating that the observed results were not particularly sensitive to demographic differences. Regression outputs can be found in Appendix C. These analyses were not preregistered.

VI. Discussion

The present study validates previous findings that suggest that transparent nudges are equally as effective as traditional nudges (Bruns & Perino, 2019). The lack of a significant difference between donation rates among transparent and traditional nudge groups indicates that transparent nudges may not be as fragile as anticipated, even in this controversial context. This is a valuable result from an ethical perspective; institutions who are not perceived as wholly prosocial can use transparent nudges without incurring a major cost.

In line with Michaelsen's previous findings, those who were in the transparent nudge condition had significantly lower perceived freedom scores than those who were in the traditional nudge condition (Michaelsen et al. 2020). This result speaks to the greater tension at play when analyzing the ethics of transparent nudging. When using a transparent nudge, the choice architect is aiming to reduce paternalism by giving users more information so that they can make a deliberate decision. The act of doing so makes users aware of the nudge, and therefore recognize that they are being influenced. This causes them to then report lower perceived freedom scores than those who were not explicitly made aware of the nudge. This is a real ethical tradeoff that should be considered when deciding whether to use a transparent nudge.

As predicted, agreement type did indeed have a large impact on donation rates. Those who agreed with the choice architect donated significantly more than both those who disagreed and those who were neutral. Perhaps the most surprising result of this study was how large of an impact neutrality had on donation rates. Those who were neutral donated far less than everyone else, with the lowest possible median donation rate of 0 and large effect sizes across comparisons. A possible explanation for this could be that neutrality was a sign of apathy; those who were neutral were not motivated by the goals of the choice architect whatsoever, and preferred to keep the tokens for themselves. Whereas those who disagreed may not have liked the Salvation Army's specific goals, but were motivated by charity generally. It also shows that those who were neutral were the least susceptible to being nudged. A further investigation into how agreement tracks with donations in other contexts would be helpful for understanding this result.

Agreement type also made an impact on perceived freedom scores. While there was no significant difference between those who agreed and disagreed, those who were neutral had uniquely different results with much higher perceived freedom scores than everyone else. This makes intuitive sense; those who were neutral donated much less, which means they were not influenced much by the nudge, and therefore did not feel any threat to their freedom. In an exploratory analysis, I looked into the differences in perceived freedom scores between transparent and traditional nudge conditions among those who were neutral (Appendix B, Figure

6). I found that those in the transparent condition felt significantly less free than those in the traditional condition, further validating the overall result. This novel finding opens up a plethora of new questions about the relationship between neutrality and behavior change interventions..

This study is just the first step in examining how transparent nudges function for controversial choice architects. In order to understand the full picture, it is imperative to test a variety of behavioral intervention types, across different contexts.

Ultimately, this study shows that nudge and agreement types have the potential to be strong mediators of both donation rates and freedom scores when analyzing behavior change interventions. This empirical information is key to understanding ethical implications relevant for choice architects, especially for those taking a subjectivist position. Continued focus on perceived freedom scores will help shed further light on this issue.

References

Bruns, H., & Perino, G. (2019). The role of autonomy and reactance for nudging-experimentally comparing defaults to recommendations and mandates. *Available at SSRN 3442465*.

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

Guala, F., & Mittone, L. (2015). A political justification of nudging. *Review of philosophy and psychology*, *6*(3), 385-395.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55-66.

Hagman, W., Andersson, D., Västfjäll, D., & Tinghög, G. (2015). Public views on policies involving nudges. *Review of philosophy and psychology*, *6*(3), 439-453

Hausman, D. M., & Welch, B. (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy*, *18*(1), 123-136.

Ivanković, V., & Engelen, B. (2019). Nudging, transparency, and watchfulness. *Social Theory and Practice*, 43-73.

Krijnen, J. M., Tannenbaum, D., & Fox, C. R. (2017). Choice architecture 2.0: Behavioral policy as an implicit social interaction. *Behavioral Science & Policy*, *3*(2), i-18.

Michaelsen, P., Johansson, L. O., & Hedesström, M. (2021). Experiencing default nudges: autonomy, manipulation, and choice-satisfaction as judged by people themselves. *Behavioural Public Policy*, 1-22.

Michaelsen, P., Nyström, L., Luke, T. J., Johansson, L. O., & Hedesström, M. (2020). Are Default Nudges Deemed Fairer When They Are More Transparent? People's Judgments Depend on the Circumstances of the Evaluation.

Schmidt, A. T., & Engelen, B. (2020). The ethics of nudging: An overview. *Philosophy Compass*, *15*(4), e12658.

Sunstein, C. R. (2017). Nudges that fail. *Behavioural public policy*, *1*(1), 4-25.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American economic review*, *93*(2), 175-179.

Appendices

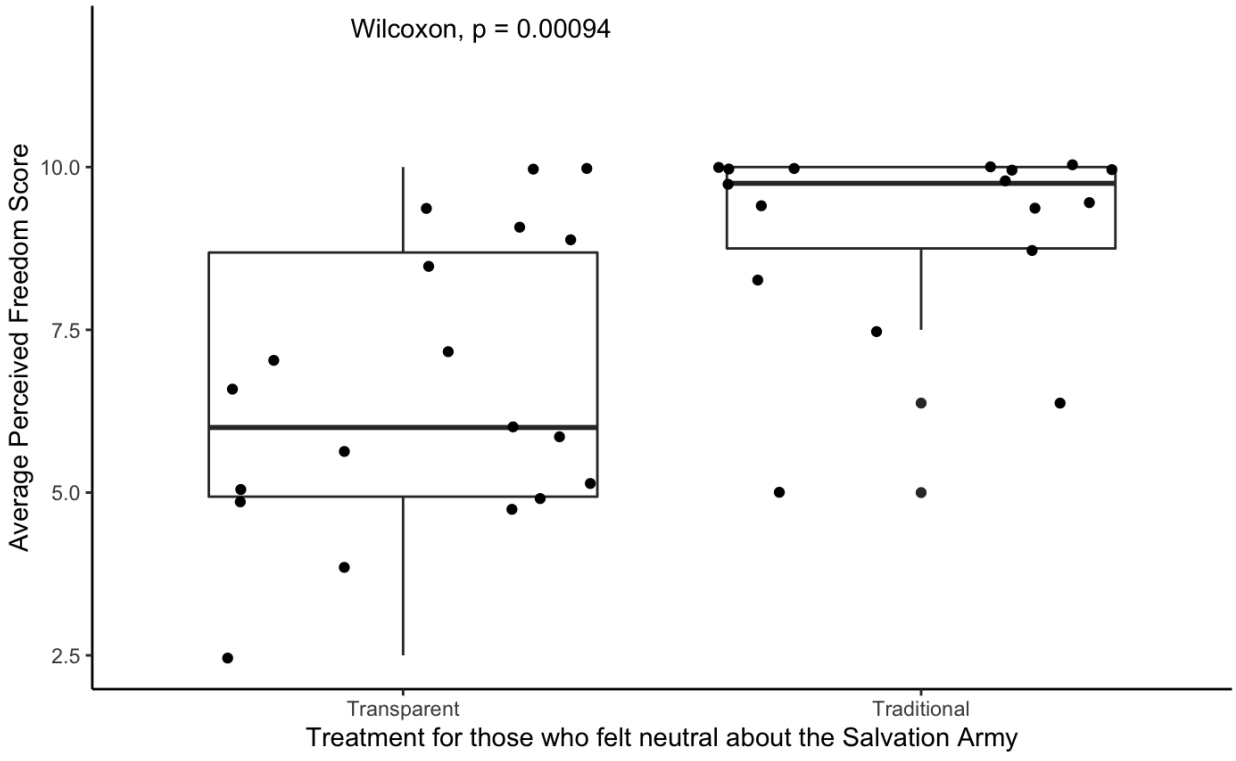
Appendix A

Salvation Army Background Statement as Seen by Participants:

- The Salvation Army "is an evangelical part of the universal Christian Church. Its message is based on the Bible. Its ministry is motivated by the love of God. Its mission is to preach the gospel of Jesus Christ and to meet human needs in His name without discrimination."
- The charity helps a variety of communities in need and is widely considered to be effective.
- Despite their current platform of inclusivity, they also have an alleged history of discrimination.

Appendix B

Figure 6 shows that among those who were neutral, those in the transparent nudge condition (M = 6) reported significantly lower perceived freedom scores than those in the traditional nudge condition (M = 9.75).



P values are from Wilcoxon rank sum tests.

Figure 6.

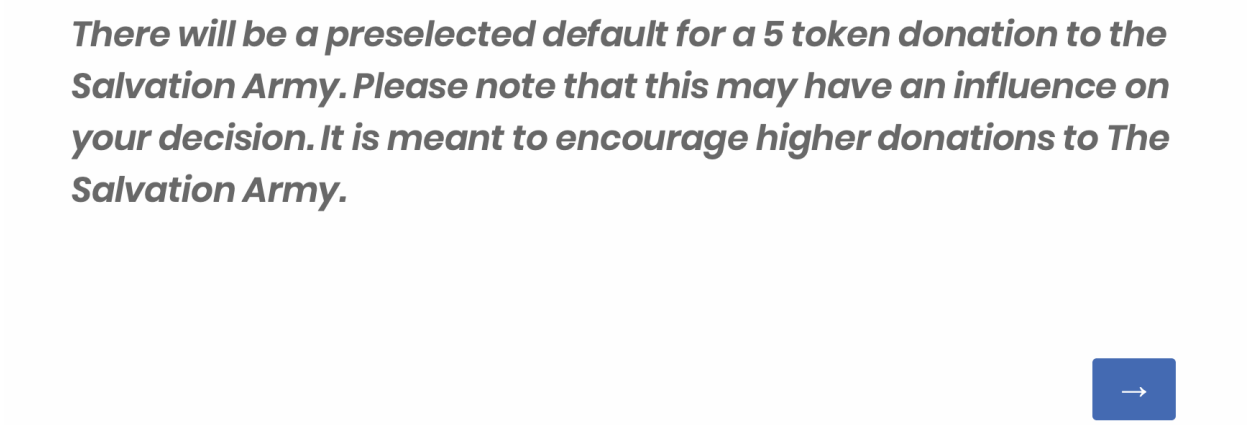
Appendix C (Regression analysis)

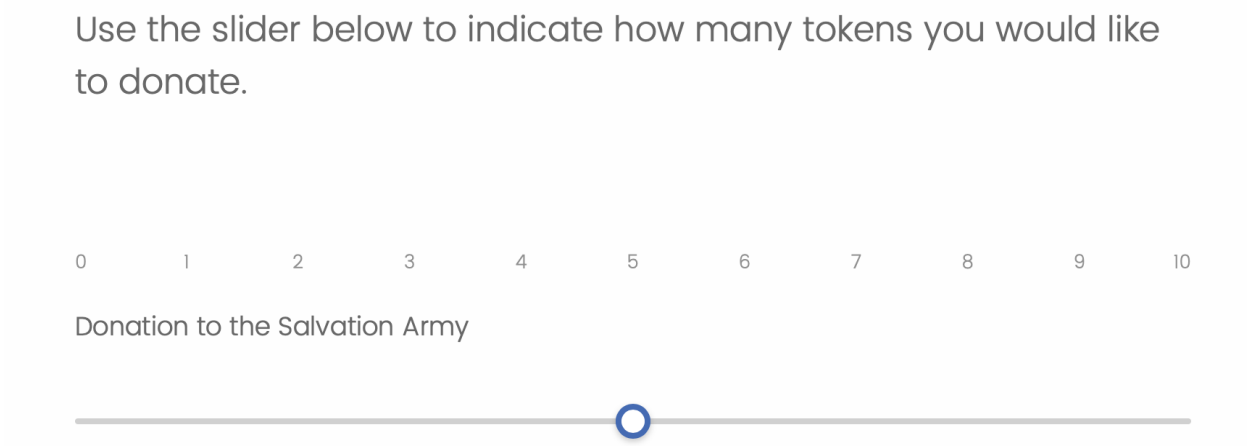
	Donation Rate by Treatment (1)	Freedom by Treatment (2)	Donation by Agreement (3)	Freedom by Agreement (4)
Transparent Condition	2.714***	-0.424		
	(0.453)	(0.267)		
Traditional Condition	2.879***	0.192		
	(0.451)	(0.266)		
Agreement(ref = Agree)				
Disagree			-3.778***	2.132***
			(0.704)	(0.512)
Neutral			-2.648***	0.601
			(0.575)	(0.418)
Gender (ref = Male): Female	0.186	-0.083	-0.100	-0.394
	(0.378)	(0.222)	(0.436)	(0.317)
Race (ref = White): Asian/Pacific Islander	1.497**	-0.191	1.200	-0.524
	(0.649)	(0.383)	(0.919)	(0.668)
Black/African American	-0.556	0.144	-1.257	0.085
	(0.728)	(0.429)	(0.898)	(0.653)
Native American	1.836	-1.039	2.340*	-2.498***
	(1.146)	(0.675)	(1.233)	(0.896)
Hispanic/Latino	-0.659	0.044	2.146*	-1.827**
	(1.162)	(0.685)	(1.250)	(0.909)
Age	-0.015	0.032***	0.028	0.020
	(0.018)	(0.011)	(0.020)	(0.014)
Constant	2.535***	5.880***	4.544***	5.726***
	(0.529)	(0.312)	(0.650)	(0.473)
Observations	304	304	161	161
R^2	0.180	0.063	0.205	0.191
Adjusted R^2	0.158	0.037	0.163	0.149
Residual Std. Error	3.158 (df = 295)	1.861 (df = 295)	2.634 (df = 152)	1.915 (df = 152)
F Statistic	8.088*** (df = 8; 295)	2.465** (df = 8; 295)	4.893*** (df = 8; 152)	4.490*** (df = 8; 152)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Appendix D (Experimental Screenshots)

Below are screenshots of the transparent nudge, and donation slider. The transparency statement was only shown to the transparent nudge condition. The slider was identical for both the transparent and traditional nudge conditions.





Appendix E (Perceived Freedom Questions)

Below are all of the perceived freedom questions that were used to calculate perceived freedom scores. All questions used 0 - 10 point Likert scales to assess agreement.

1. To what extent do you feel in control of your donation decision?
2. To what extent do you feel your donation decision was thought through?
3. To what extent do you feel your donation decision reflected your preferences?
4. To what extent do you feel that your donation decision was free from external influence?

Please indicate the degree in which you agree with the following statements:

"The way in which the donation choice was presented to me…"

1. "... threatened my freedom to choose what I wanted."
2. "... made me feel like my decision was made for me."
3. "... tried to manipulate me."
4. "... tried to pressure me."