Do I Trust this Stranger? Generalized Trust and the Governance of Online Communities

Jérôme Hergueux
French National Center For Scientific Research
(CNRS, BETA lab)
Berkman Klein Center for Internet & Society at
Harvard University
FRANCE
jerome.hergueux@gess.ethz.ch

Yann Algan
Sciences Po, Department of Economics
FRANCE
yann.algan@sciencespo.fr

Yochai Benkler
Harvard Law School
Berkman Klein Center for Internet & Society at
Harvard University
USA
ybenkler@law.harvard.edu

Mayo Fuster-Morell
Open University of Catalonia
Berkman Klein Center for Internet & Society at
Harvard University
SPAIN
mfuster@uoc.eu

ABSTRACT

Online peer production communities such as Wikipedia typically rely on a distinct class of users, called administrators, to enforce cooperation when good faith collaboration fails. Assessing one's intentions is a complex task, however, especially when operating under time-pressure with a limited number of (costly to collect) cues. In such situations, individuals typically rely on simplifying heuristics to make decisions, at the cost of precision. In this paper, we hypothesize that administrators' community governance policy might be influenced by general trust attitudes acquired mostly out of the Wikipedia context. We use a decontextualized online experiment to elicit levels of trust in strangers in a sample of 58 English Wikipedia administrators. We show that low-trusting admins exercise their policing rights significantly more (e.g., block about 81% more users than high trusting types on average). We conclude that efficiency gains might be reaped from the further development of tools aimed at inferring users' intentions from digital trace data.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**;

KEYWORDS

Online Experiment, Trust, Community Policing, Peer Production, Wikipedia

ACM Reference Format:
Jérôme Hergueux, Yann Algan, Yochai Benkler, and Mayo Fuster-Morell. 2021. Do I Trust this Stranger? Generalized Trust and the Governance of Online Communities. In *Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3442442.3452338

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452338

1 INTRODUCTION

Online peer production communities developing global public goods such as Wikipedia depend on their ability to attract and retain volunteer contributors to thrive. Over the past decade, however, the encyclopedia that "anyone can edit" has experienced difficulties in expanding its contributor base. The increasing bureaucratization of Wikipedia [2], together with the impersonal enforcement of an ever-growing number of policies aimed at protecting the encyclopedia from low quality and bad faith contributions, has been pointed out as one of the major factors behind its current difficulties at retaining editors [3, 10, 12].

As Wikipedia remains committed to openness, a relatively small number of administrators face the daunting task of trying to strike the right balance between "exclusion" (of malicious actors) and "inclusion" (of good faith contributors). Wikipedia administrators are editors who successfully self-selected into a highly competitive peer-review process, at the end of which they were entitled with special oversight rights over the community. These editors are notably in charge of blocking disruptive users, deleting pages that they consider will not develop as proper encyclopedic articles and protecting vandalized pages from being edited. Overall, there are about 1,400 contributors holding admin rights on English Wikipedia.

Each day, administrators need to deal with a large number of potentially malicious users and damaging contributions. They need to make quick inferences as to whether individuals are acting in good faith so as to prevent threats, protect the common resource and enforce cooperation whenever necessary. The key challenge that administrators face is a complex one. They need to jointly minimize the risk of (i) failing to detect harmful behavior, and (ii) wrongly exercise their policing rights on well-intentioned editors, potentially driving them away from the community.

As experienced as they may be, administrators operate under both time-pressure and uncertainty. They therefore likely rely on a

number of (possibly) inefficient decision-making heuristics to ease their decision making process [7, 15]. Accordingly, a number of researchers have started to develop classifiers and tools to guide them in those complex decisions [9, 11].

The goal of this paper is to provide a direct test of the hypothesis that Wikipedia administrators rely on such decision making heuristics when making their community policing decision. More specifically, we focus our analysis on one prominent social heuristics: the level of trust in strangers, or "generalized trust" [5]. Trust is a behavioral heuristic that people develop early in their social lives in the context of the institutions and social norms that govern their daily interactions. In the face of uncertainty about the intentions of (possibly anonymous) editors, we hypothesize that general trust attitudes acquired mostly out of the Wikipedia context might influence the governance policy of its administrators.

To test this hypothesis, we run an online experiment on the English Wikipedia website, and show that administrators' level of trust in strangers significantly predicts the use of their community policing rights.[1] For instance, controlling for a vector of demographic variables, we find that low trusting administrators block about 81% more users than high trusting ones. We conclude that efficiency gains might be reaped from the further development of tools and classifiers aimed at inferring users' intentions from digital trace data, so as to better guide administrators in their community policing decisions.

2 THE ONLINE EXPERIMENT

Complex social constructs such as trust are difficult to measure with traditional survey tools [8]. To elicit Wikipedia administrators' general trust attitudes, we thus rely on a decontextualized experimental game used to elicit trust attitudes in the laboratory: the Trust game [1]. We use the Wikimedia banner system as a convenient recruitment device for our online experiment. The Wikimedia Foundation relies on this banner system to advertise its annual fundraising. It is also used by the community of editors for purposes of extended internal communication (e.g., to advertise events and other community initiatives). As a result, the banner system represents a powerful and trusted way of reaching out to Wikipedia editors. In coordination with the Wikimedia Foundation, we coded this recruitment banner so that it would be displayed at the top of every Wikipedia page for all registered users eligible to participate in the experiment[2], until he or she decided either to click on it, or to disable it. Figure 1 features this recruitment banner.

Upon clicking on the banner, subjects were uniquely and automatically identified through their Wikipedia user id number and redirected to the welcome screen of our experimental economics platform. The welcome page of the decision interface provided subjects with general information about the experiment, including the number of sections, the expected completion time, and how their earnings would be computed and paid through Paypal. In order to minimize potential demand effects and in-group biases, we took great care to present the study as non Wikipedia oriented, and made it very clear on the introductory screen that subjects would interact with a diverse pool of Internet users. (See the Appendix for a detailed account of our experimental procedures.) The experiment was launched on December 8th 2011 and the banner recruited 120 Wikipedia administrators in 8 hours (representing about 9% of the overall admin population).

The Trust game involves a one-shot anonymous interaction between two individuals, a trustor and a trustee. Depending on their login order, subjects were sequentially attributed one of those roles: either participant A (the trustor) or participant B (the trustee). At the beginning of the game, both players receive a $10 initial endowment. The trustor then faces the following decision: she has the opportunity to transfer any amount taken from her endowment to the trustee. (Any amount not transferred is kept by the trustor with certainty.) The trustee receives three times the amount sent by the trustor, and has to decide how much to send back, if anything.[3] (See the instructions screen of the Trust game in Figure 2.)

Efficiency is achieved in this game when the trustor sends her full $10 endowment to the trustee, who then receives $30. But making this decision requires that the trustor places significant trust on her anonymous counterpart, as the trustee cannot commit nor be forced to send anything back. As a result, the amount (from 0 to 10) sent by the trustor can be interpreted as a measure of trust in strangers, or generalized trust [1, 5, 8].[4]

At the end of the experiment, we asked subjects some standard demographic information questions. Since the decision to trust in the experiment might also be correlated with subjects' attitudes towards risk (and despite empirical evidence to the contrary [14]), we also ask subjects an experimentally validated question on risk aversion [4].

3 DATA AND VARIABLES

Out of the pool of 120 admin participants to this experiment, 58 were assigned the role of trustor. For those subjects, we use the proportion of their endowment which they decided to send to the anonymous trustee as our experimental measure of generalized trust.

Table 1 summarizes our explanatory variables, i.e., subjects' level of trust in strangers, age, gender, degree level[5] and risk attitudes.[6] We can see that our subjects are relatively young on average (32 years old), predominantly male (90%), and typically have some college education. On average, subjects decided to send 68% of their endowment to trustees, with significant heterogeneity across players.

In order to capture administrators' community policing decisions, we collect for each subject: (i) the total number of users blocked, (ii)

[1] We focus on the English Wikipedia because it can be safely assumed to be the most "mature" Wikipedia community. Still, numerous independent language editions of Wikipedia exist, where it may be interesting to replicate these findings.

[2] On top of Wikipedia administrators, the banner also targeted regular contributors to play a Public Goods game. We do not analyse those decisions in the context of this paper.

[3] We elicit this decision through the strategy method: for each possible transfer from the trustor (from 1 to 10) the trustee chooses how much will be returned without knowing the trustor's actual choice.

[4] Note that *both* players receive the same $10 endowment at the beginning of the game, so that fairness preferences cannot interfere with the trustor's decision.

[5] This variable was reported on a 8-points scale: 1 = "less than high school"; 2 = "high school"; 3 = "some college"; 4 = "2 years college degree"; 5 = "4 years college degree (BA, BS)"; 6 = "masters degree"; 7 = "professional degree (MD, JD)"; 8 = "doctoral degree".

[6] On a 10-points scale, subjects are asked whether they generally see themselves as fully prepared to take risks as opposed to generally trying to avoid taking risks: 0 = "unwilling to take risks" and 10 = "fully prepared to take risks".

Figure 1: The Wikipedia recruitment banner

Figure 2: The Trust game instruction screen

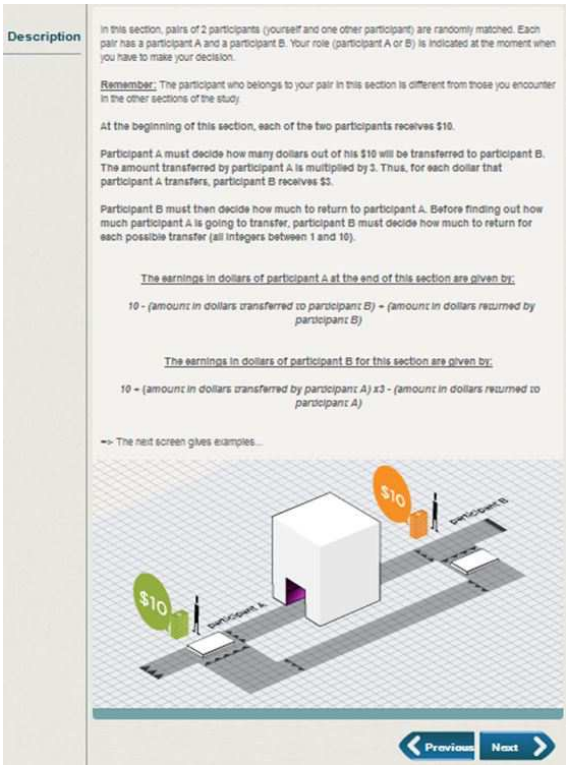

Table 1: Explanatory variables: online experiment

	Obs	Mean	Std. Dev.	Min	Max
1. trust	58	0.68	0.29	0	1
2. age	58	32.26	10.71	18	69
3. female	58	0.10	0.31	0	1
4. degree level	58	4.86	1.72	2	8
5. risk aversion	58	5.72	2.43	0	10

the total number of pages deleted, (iii) the total number of paged protected from editing, and (iv) the overall number of admin actions performed. We do this starting from the time of the experiment (i.e., December 8, 2011), and up to October 6, 2020 (i.e., for about 9 consecutive years). As an additional source of data, we also got back in touch with our subjects 6 months after the completion of the experiment (i.e., in July 2012), and asked them to estimate the fraction of their working time on Wikipedia that they had spent on policing tasks (e.g. deleting and protecting pages, blocking and unblocking users etc.), as opposed to regular contribution activities (on a scale from 1 to 10). We received 27 answers out of our initial pool of 58 subjects. Table 2 summarises our dependent variables.

Table 2: Dependent variables: community policing

	Obs	Mean	Std. Dev.	Min	Max
1. user block	58	2422.07	8212.77	2	57624
2. page delete	58	5924.52	15215.28	11	109653
3. page protect	58	479.40	1326.41	0	7477
4. admin actions	58	11693.47	25604.89	152	183935
5. admin time	27	4.19	2.62	1	10

One potential concern with the above variables is that they might be correlated with administrators' level of activity as regular contributors. As a result, one might erroneously conclude that trust is related to community policing, while it may in fact simply be related to overall wiki activity. In order to address this concern, we collect one more control variable on top of those presented in Table 1: the number of edits performed in the English Wikipedia main namespace (i.e., manual edits to Wikipedia articles).[7] By adding this control, we effectively estimate the relationship between trust and community policing while holding regular editing activity constant. (This variable is distributed with a mean of 9,597, a standard deviation of 19,956, and a min and max value of 17 and 101,208, respectively.)

4 RESULTS

Table 3 presents OLS estimates (with robust standard errors) of the effect of generalized trust on the community policing activity patterns of Wikipedia administrators. Since the distribution of variables 1-4 in Table 2 is heavily skewed to the left (i.e., follows a power-law distribution), we log-linearize those dependent variables in Table 3. For the same reason, we also log-linearize the number of mainspace edits in the control variables.

Some focus on our control variables indicates that age is negatively associated with the number of users blocked (i.e., 4.3% decrease with each additional year on average, column (1)). For some

[7] By "manual edits" we mean that we exclude from this count the edits that administrators perform with the help of semi-automated tools, which are typically used to perform maintenance tasks such as fighting vandalism (see [6].)

reason, female administrators also seem to engage less in community policing (e.g., 70% decrease in the number of admin edits, column (4)). Education appears to be negatively related with community policing, but we lack statistical power to estimate those coefficients precisely. Risk aversion is significantly associated with the fraction of their editing time that administrators declared spending on policing tasks (column (5)). Finally, the level of regular editing activity is positively related to the number of pages protected (column (3)) and the overall number of admin actions performed (column (4)), but is negatively related to the time spent on admin activities (column (5)).

Turning our attention to our coefficients of interest, we find that moving from no trust to full trust in strangers is associated with a 81% decrease in the total number of users blocked from editing (column (1)).[8] It is also associated with a 81% decrease in the number of pages deleted (column (2)), a 63% decrease in the number of pages protected from editing (column (3), although this coefficient does not reach statistical significance), and a 70% overall decrease in the number of admin tasks performed (column (4)). Finally, column (5) of Table 3 presents the OLS estimate of the relationship between trust in strangers and the fraction of their working time on Wikipedia that administrators report dedicating to policing activities. Despite the small sample size, moving from no trust to full trust in the experiment is significantly associated with a 4.28 points decrease in the proportion of editing time dedicated to admin activities (out of a 10 points scale, this corresponds to a decrease of 1.55 standard deviations).

5 CONCLUSION

Online peer production communities such as Wikipedia typically rely on an elevated category of users, called administrators, to enforce cooperation whenever good faith collaboration fails. Assessing editors' intentions is a complex task, however, especially when an important feature of administrators' role is to operate under both time-pressure and uncertainty. In practice, administrators often need to rely on a limited number of (costly to collect) cues when deciding whether to assume good faith or take action to protect the common resource. In such complex decision making environments, individuals typically rely on simple heuristics to ease their decision making process, at the cost of precision [7, 15]. The complex question of "should I block this new user?" can then be substituted for an easier one: "do I trust this stranger?".

In this paper, we focus on trust in strangers as a social heuristic especially relevant to community policing in peer production communities committed to openness such as Wikipedia. We rely on a decontextualized online experiment to elicit Wikipedia administrators' general trust attitudes, and use this measure to provide a direct test of the link between trust attitudes and community policing. We find that Wikipedia administrators heavily rely on trust as a social heuristic to guide their community governance decisions. As other researchers have already argued in other contexts [9–12], this result suggests that efficiency gains could be reaped from an investment into developing classifiers and tools aimed at helping administrators infer contributors' intentions (or "good faith") from their activity traces [3, 16].

[8]The exact effect size is computed as $e^{\hat{\beta}} - 1$.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the ERC Starting grant, the University of Strasbourg Attractivity grant, and logistical support from the Sciences Po médialab and the Wikimedia Foundation. We are thankful to Anne l'Hôte, Romain Guillebert, David Laniado and Tarun Chadha for outsdanding research assistance. We are indebted to Dario Taraborelli for his help in moving this project forward and Timo Tijhof for coding the recruitment banner. Last thanks go to the Wikipedia administrators who took part in this study.

REFERENCES

[1] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 1 (1995), 122–142.
[2] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1101–1110.
[3] Jonathan Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of blocked community members: Redemption, recidivism and departure. In *The World Wide Web Conference*. 184–195.
[4] Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association* 9, 3 (2011), 522–550.
[5] Ernst Fehr. 2009. On the economics and biology of trust. *Journal of the european economic association* 7, 2-3 (2009), 235–266.
[6] R Stuart Geiger and David Ribes. 2011. Trace ethnography: Following coordination through documentary practices. In *2011 44th Hawaii international conference on system sciences*. IEEE, 1–10.
[7] Thomas Gilovich, Dale Griffin, and Daniel Kahneman. 2002. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.
[8] Edward L Glaeser, David I Laibson, Jose A Scheinkman, and Christine L Soutter. 2000. Measuring trust. *The quarterly journal of economics* 115, 3 (2000), 811–846.
[9] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.
[10] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist* 57, 5 (2013), 664–688.
[11] Aaron Halfaker, R Stuart Geiger, and Loren G Terveen. 2014. Snuggle: Designing for efficient socialization and ideological critique. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 311–320.
[12] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*. 163–172.
[13] Jérôme Hergueux and Nicolas Jacquemet. 2015. Social preferences in the online laboratory: a randomized experiment. *Experimental Economics* 18, 2 (2015), 251–283.
[14] Daniel Houser, Daniel Schunk, and Joachim Winter. 2010. Distinguishing trust from risk: An anatomy of the investment game. *Journal of economic behavior & organization* 74, 1-2 (2010), 72–81.
[15] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
[16] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. 2020. The effects of algorithmic flagging on fairness: quasi-experimental evidence from Wikipedia. *arXiv preprint arXiv:2006.03121* (2020).

APPENDIX: EXPERIMENTAL PROCEDURES

Our design strictly follows the experimental procedures detailed in [13]. Those procedures have been developed specifically so as to strengthen the internal validity of Internet-based experiments. Their reliability was established through a careful comparison of decisions elicited in the lab and online.

In order to minimize potential demand effects and in-group biases when eliciting subjects' trust attitudes, we take great care to present the study as non Wikipedia oriented. We make it very clear on the introductory screen that subjects will interact with a

Table 3: Trust and community policing

	(1)	(2)	(3)	(4)	(5)
	log(user block)	*log*(page delete)	*log*(page protect)	*log*(admin actions)	admin time
trust	-1.668**	-1.654*	-0.989	-1.212*	-4.278***
	(0.822)	(0.847)	(0.767)	(0.637)	(1.497)
age	0.0434*	0.0348	0.0205	0.0236	-0.00723
	(0.0252)	(0.0236)	(0.0231)	(0.0164)	(0.0402)
female	-0.861	-0.555	-1.203**	-0.706*	-0.510
	(0.860)	(0.416)	(0.591)	(0.358)	(1.538)
degree level	0.00271	-0.100	-0.176	-0.0862	-0.375
	(0.182)	(0.155)	(0.163)	(0.128)	(0.273)
risk aversion	-0.0502	-0.0384	-0.00773	-0.000281	0.498***
	(0.116)	(0.0837)	(0.0993)	(0.0673)	(0.133)
log(mainspace edits)	0.140	0.127	0.272**	0.217**	-0.445***
	(0.155)	(0.107)	(0.112)	(0.0906)	(0.139)
Constant	4.778***	7.235***	3.517***	7.342***	10.13***
	(1.530)	(1.185)	(1.220)	(0.938)	(2.318)
N	58	58	58	58	27
adj. R^2	0.0404	0.0540	0.128	0.164	0.347

diverse pool of Internet users.[9] Subjects were only informed of their earnings at the very end of the experiment. In addition to those earnings, subjects were guaranteed to receive a $10 participation fee. Subjects got paid upon completion of the experiment through an automated PayPal transfer.[10] We only required a valid e-mail address to process the payment. To strengthen the credibility of the payment procedure, we asked subjects to enter the e-mail address that is (or would be) associated with their PayPal account right after the introductory screen of the decision interface.

It is important to stress that Wikipedia contributors can be very hostile to monetary rewards. In order to ensure that the experiment was equally incentive compatible for all subjects, we allowed them to donate any amount taken from their final earnings to the Wikimedia Foundation and/or the International Committee of the Red Cross, a renowned and general purpose charitable organization, in anticipation of the fact that some subjects might not want to donate to the Wikimedia Foundation. This possibility was made clear on the welcome screen of the decision interface. It was not possible, however, to commit to donating one's final earnings prior to the study's completion. In all cases, subjects' decisions were made under full anonymity.

One important methodological concern with the online implementation of the experiment is to guarantee a quick and appropriate understanding of the instructions when no interaction with the experimenter is possible. We strengthened the internal validity of the online experiment with three distinctive features of the interface. First, we included suggestive flash animations illustrating the written experimental instructions at the bottom of the instruction screen (see Figure 2).[11]

Second, the instructions were followed by a screen providing some examples of decisions, along with the detailed calculation of the resulting payoffs for each player. These examples were supplemented on the subsequent screen by an earnings calculator. On this interactive page, subjects were allowed to test all the hypothetical scenarios they were interested in before making their decisions. In contrast to the illustrative flash animations, the numeric results of each scenario run by a subject in the earnings calculator screens were explicitly displayed.

Last, the system provided a quick access to the instructions material at any moment during decision-making. On all screens, including decision-making ones, a "review description" button gave subjects a direct access to the instructions displayed at the beginning of the game. The system also allowed participants to navigate at will from one screen to another through the "Previous" and "Next" buttons located at the bottom of each screen.

[9] In practice, subjects' decisions were randomly matched with a pool of decisions elicited from their fellow Wikipedia editors, online students and open source software developers.

[10] Such a payment procedure guarantees a fungibility similar to that of cash transfers in lab experiments, as money transferred via PayPal can be readily used for online purchases or easily transferred to one's personal bank account at no cost.

[11] The loop of concrete examples displayed in each animation was first randomly determined and then fixed for each game. The same loop is displayed to all subjects without any other numeric information than the subjects' initial endowments.