Standing naked in the storm– European citizens' trust in social media, users, information

Bodo, B.; Bene, M.; Boda, Z.

DOI
10.2139/ssrn.4368419

Publication date
2023

Document Version
Final published version

UNIVERSITY OF AMSTERDAM

IViR

AMSTERDAM CENTER FOR LAW & ECONOMICS

STANDING NAKED IN THE STORM– EUROPEAN CITIZENS' TRUST IN SOCIAL MEDIA, USERS, INFORMATION

Balazs Bodo

Márton Bene

Zsolt Boda

Amsterdam Law School Legal Studies Research Paper No. 2023-12

Institute for Information Law Research Paper No. 2023-02

Amsterdam Center for Law & Economics Working Paper No. 2023-04

Standing naked in the storm– European citizens' trust in social media, users, information

Bodo, Balazs[1]; Bene, Marton[2, 3]; Boda, Zsolt[2,]

Abstract

We have surveyed users in 7 European countries about the factors that underwrite their trust in and on Facebook. We identified three trust pillars: (1) self-confidence to recognize and hedge platform related risks; (2) trust in the platform's ability and willingness to protect users against online harms; and (3) national and European platform regulation, and measured their role in shaping users' trusting attitudes towards the platform and the users and information they encounter on the platform.

Our finding suggests serious deficiencies in how trust in such a critical societal infrastructure is structured. First, we found that trust in and on the platform is mostly defined by generic, rather than platform-specific trust attitudes and expectations. Second, our data suggests that the two most important platform-specific trust pillars are rather shaky. On the one hand, users trust the platform to protect them, while there is ample evidence of Facebook and its parent company, Meta to not act in the best interest of their users. On the other hand, users base their trust on their perceived ability to protect themselves, while our data shows that they do not seem to use the even the very limited set of tools the platform provides them for such self-protection.

Lastly, while governments seem to be best positioned to reign in global digital service providers, and in recent years the EU spent enormous amounts of resources to regulate online platforms, users don't seem to expect regulation to make platforms and the information herein trustworthy. Such lack of expectations towards the only agents in this ecosystem who can provide any effective trustworthiness safeguards vis-à-vis platforms is somewhat disturbing.

[1] Institute for information Law, University of Amsterdam.

[2] Center for Social Sciences, Hungarian Academy of Sciences ; ELTE Eötvös Loránd University, Budapest, Hungary

Introduction

The Fall of 2022 will go down in the history books as the parallel fall of crypto and Twitter. The double crisis of a centralized and a decentralized techno-social system points to fundamental questions around the trustworthiness of critical infrastructures in the digital society.

Blockchain technology was invented in response to the mistrust in banks and governments after the 2008 financial crisis. It promised that one does not need to trust traditional institutions, nor there is a need for regulation, because a trustworthy system of value exchange can be built solely on cryptography, transparent code, cryptoeconomic incentives, rules coded into, and self-executed by the infrastructure. 14 years after its invention, in the matter of a few short months institutional and retail investors lost more than two trillion dollars, because the "trust machine" (Anon 2015) turned out to be an untrustworthy mess of Ponzi-schemes, and other fraud. "Lex cryptographia" (Wright and De Filippi 2018), the idea that code can be trustworthy in and by itself, without regulation, institutional control and oversight, turned out to be a baseless aspiration (Quintais et al. 2019).

Meanwhile, in October 2022 Elon Musk took Twitter private, and started to change the governance of the platform. The consequences were immediate: trolls and parody accounts started to impersonate public figures, companies, and brands, causing sometimes billions of dollars of damage. Key personnel, responsible for content moderation, the trust and safety of the platform also left. As a result, millions of users left the platform, as did top advertisers. (Kann and Carusone 2022; Stokel-Walker 2022)

The troubles at Twitter and in the crypto sphere are just the tip of the iceberg. They point to a fundamental problem in digital society: we extensively rely on private technological infrastructures to engage in trust-requiring social, economic relations, but these infrastructures are often untrustworthy. The extent to which users will trust a seller/buyer on an electronic marketplace, feel safe to sit in a stranger's car, sleep in a stranger's bed, or believe a piece of information coming from a stranger will (or at least should) depend on the trustworthiness, i.e., the benevolence and ability of eBay, Uber, Airbnb or social media platforms to identify and ban fraudsters, misinformation and other untrustworthy elements on their services.

This rise of technology-mediated trust (Bodó 2020), and the new trust infrastructures (Bodó 2021) impact all forms of trust relations in society. Oftentimes technology developers want and succeed to disrupt well-established interpersonal and institutional ways to maintain trust in society. Almost all of our trust relations are affected by this dynamic, from intimacy, disrupted by online dating (Bonilla-Zorita, Griffiths, and Kuss 2020; Ortega and Hergovich 2018), via AI systems upsetting the democratic role of media (Helberger 2019), to social cohesion, disrupted by predictive policing, or automated welfare systems (Bodó and Janssen 2022). In response, to the widespread use, the apparent trustworthiness of these systems the European Union introduced new pieces of legislation such as the Digital Services Act, the AI Act, or the Markets in Crypto Assets Regulation to address trust concerns in the digital society. The new rules and institutions are designed to identify the individual and societal risks and harms and force technology developers and operators to take the interests of these stakeholders seriously. But role states play in the digital trust domain is more than providing trustworthiness safeguards through regulation, and public institutions (Sztompka 1999). Both autocratic and democratic states were caught increasing societal mistrust, by, for example, using social media to upset the political stability abroad. (Keller and Klinger 2019; Spaiser et al. 2017)

In short, our interpersonal relations, institutions, and societies are penetrated by techno-social systems which produce new forms of trust relations, remediate, and disrupt existing ones. These

developments lead to the following questions: what kind of trust these technical infrastructures produce? To what extent are these infrastructures trustworthy? To what extent they are trusted? And what are the sources of trust in and by these systems?

In this article we take the first, explorative steps to empirically describe the patterns of trust in the digital society. We survey European citizens' attitudes in technology-mediated trust relations. We focus on Facebook to understand the risks users associate with the use of the platform; and how trust is used to overcome those risks. Our research measures the role of three pillars on which users' trust is resting: users' self-confidence in detecting and counter risks; trust in the platform's willingness and ability to protect users from harms; and expectations vis-à-vis the government to regulate social media. By describing the role of these trust pillars, we identify the gaps in the current approaches of creating a more trustworthy digital society and identify new opportunities to establish *trustworthy and trusted* digital infrastructures.

Trust and trustworthiness in the digital context - literature review

Trust as an attitude

Trust is a widely studied phenomenon. It is believed to be both the indicator and the cause of many social phenomena with a positive valence. Trust promotes cooperation between people and with institutions; it predicts, among others, compliance with the law, and with health measures during the pandemic, willingness to pay taxes, and acceptance of new technologies (Bargain and Aminjonov 2020; Khan 2022; Murphy 2005). Trust is also an elusive concept. Its varied and context-dependent roots make it difficult to trace the factors behind it, even though it would be important to identify its causes given its many alleged benefits. Below we shortly summarize the most important findings in this regard, focusing on the problem of trustworthiness.

Trust is defined by psychologists as an attitude held by individuals: it is a future-oriented expectation that the interaction with the other (be it another person, a group of persons, a technology or an institution) will produce positive outcomes for the trustor (Hardin 2002). Therefore, it is only meaningful to talk about trust if at least some risks are at stake (Luhmann 2017). Since the very concept of risk implies a level of calculability, some argue that trust is, or should be, a rational phenomenon building on past experiences, reasonable expectations, and external factors to control risks (e.g., regulation, laws, enforced social norms) (Hardin 2002; Norris 2022). However, trust has proven to be a much more complex phenomenon, dependent on a number of both individual level and macro level factors.

On the one hand, trust is influenced by individual-level factors like income, education, or personal characters – wealthier, more educated, left-leaning, politically more interested people and men, for instance, tend to have higher trust in both institutions and other people (Boda & Medve-Bálint, 2014; Christensen et al., 2020; Zmerli & Newton, 2011). The fact that personality traits as well as socio-demographic characteristics have a statistically significant predictive power suggests that trust is partly a non-rational, unreflective phenomenon (Lagerspetz 1998).

On the other hand individual trust is influenced by macro-level factors, like income (GDP) or social inequality – high income countries with lower level of income inequality tend to enjoy higher confidence in both people and institutions (Anderson & Singer, 2008; Medve-Bálint & Boda, 2014). Trust is also influenced by culture. Fukuyama (1995) has shown that trust is deeply rooted in different

social, economic, political histories and experiences. For example, if trust in public institutions is low, trust necessitating interactions will rely more on interpersonal trust relations: familiar networks, religious affiliations, and the like. The prominence of one trust infrastructure over the other is rooted in the given society's history. In other words, trusting attitudes, including institutional trust, are part of a larger belief-system that influences how and how much people trust each other and organizations (Giddens 1990).

Finally, trust is also shaped by the perceptions concerning the trustworthiness of the trustee.

Trust and trustworthiness

Trust is an attitude of a trustor in a trust relationship. Trustworthiness, on the other hand, is the attribute of the trustee. Trustworthiness is the perception that though the trustee has considerable freedom to cause harm to the trustor, the trustee will not abuse its agency at the detriment of the trustor. The perception of trustworthiness provides a rational foundation for trusting attitudes (Hardin 2002). Two fundamental questions arise concerning trustworthiness. First, what constitute trustworthiness; second, what shapes the perception of trustworthiness.

Trustworthiness has different conceptualizations, but they necessarily include the two basic dimensions of competence and benevolence (Levi, Sacks, and Tyler 2009; Mayer et al. 1995). First, trustworthiness needs the competence of the trustee to deliver the expected outcomes in the domain where trust is given. Second, a trustee (which can be a person, an institution, or a techno-social system) can be seen trustworthy if it is seen to be benevolent, i.e., acting in the best interest of the trustor.

The second question is what shapes the perception of trustworthiness. A rational actor would look for information on the trustor, such as past experiences or the accounts of others on the trustworthiness of the trustor. Trustworthiness, especially in the dimension of competence, may also be supported by institutional guarantees and third-party certifications: diplomas and professional agencies that assess the competences of the trustor, as well as authorities that enforce safety regulations (Hardin 2002: 9). However, the perception of trustworthiness often relies on proxy indicators – this is especially true for the benevolence dimension which is notoriously difficult to assess. One such proxy is the perception of value alignment between the trustor and trustee: if we share basic moral values with the trustee our perception of their trustworthiness is supported (Mayer et al. 1995). Another proxy is whether the behavior of the trustee conforms to the principles of procedural fairness. While the role of procedural fairness perception in building trust has been demonstrated in many different social settings (see, e.g., Hawdon 2008; Levi et al. 2009; Murphy 2005), the mechanism behind this effect is less clear. Smith et al. (2007) argue that conformity to procedural fairness is used as an indicator on the benevolence of the trustee.

Research questions and hypotheses

Bodó argues that despite the centrality of various techno-social systems in the digital society, and in our everyday lives, it is difficult, if not impossible to objectively verify these trustworthiness dimensions, and the available evidence points to them being fundamentally untrustworthy (Bodó 2020, 2021).

The fact that despite such fundamental shortcomings untrustworthy techno-social systems are still widely used points us to two possible explanations. First, users' trust in fundamentally untrustworthy systems may be a sign of unreflective trust: not realizing or disregarding potential risks (O'Neill 2002). Second, users may be well aware of the risks of the systems they use, but they trust them nevertheless, because they feel that they possess the right tools, and/or they can rely on internal and external safeguards to protect themselves from the possible risks and harms.

In our study we try to address both issues. First, we try to find evidence of unreflective trust, i.e., users using social media platforms without being aware of the risks they face. Second, we identified on what pillars users base their trust in the platform, and trust via the platform in the users and information they encounter there. Do they feel they possess knowledge and tools to protect themselves (trust in one's own competencies)? Do they feel that the platform operator would protect them (benevolence and integrity of the platform trustee)? Do they expect the authorities to step in their defense (competence, benevolence of the state)?

In accordance with the theoretical models, we ask these questions in relation to two different trustees. On the one hand, we interrogate what shapes the trustworthiness perceptions of other users, and information encountered online. On the other hand, we ask what makes the platform itself trustworthy in the eyes of its users. The two questions are closely related as the trustworthiness of the users and information on the platform partly depend on the trustworthiness of the platform, as the platform mediates trust relationships between users, and users and information (Bodó 2020).

In particular, we defined our research questions as follows.

We first posed an open research question about the risks which users in the surveyed European countries associate with using various Meta properties. We also formulated the following hypotheses:

> *H1: there is a significant correlation between the perceived risks associated with the technology, and users' trust in the platform, its users, and the information circulating on the platform.*

First, we test the naïve hypothesis, that those with higher risk perception/awareness would trust less the platform, its information, and users.

Second, we test a more complex hypothesis, which theorizes that risk-aware respondents may still demonstrate some trust in and on the system, if they feel they have some form of support to counter those risks. In particular we identified three potential pillars to support trust, which may underline the use of a system fraught with risks: users' confidence in their own powers to detect harms and defend themselves; users trust in Meta so it will do its best to protect them; and users' confidence in their national and EU governments to make platforms a safe place through regulation.

In concrete:

> *H2a: users will have more trust in the platform, the users, and the information if they have confidence in their own ability to recognize risks.*
>
> *H2b: users will have more trust in the platform, the users, and the information in they feel they can protect themselves from the harms in the system*
>
> *H3: users will have more trust in the platform, the users, and the information if they feel they can trust the platform operator to protect them from harms (through algorithmic or human policing and moderation of harmful content).*

H4: users will have more trust in the platform, the users, and the information, if they feel their national governments, or European institutions will protect them from risks and harms through regulation.

In the next step we interrogated the extent to which these different safeguards can substitute each other. Will those users who have less faith in the platforms self-regulating capacities trust their own skills to defend themselves against online risks? Will those who trust less themselves look for help from the government?

H5: there is a significant substitution effect between the different trust pillars. Lower levels of trust in one will imply higher levels of trust in the other pillars for the same levels of overall trust in the platform, the users, and the information

We also hypothesize that both the effect of the three trust pillars and the substitution effects between them will be different for users with different knowledge of and attitudes towards risks.

Literature on trust differentiates between unreflective trust based on blind faith versus calculative trust based on the careful assessment of risks and ways to mitigate them. In that framework we expect risk-conscious users to have higher importance to different trustworthiness safeguards than those, who seem to be less knowledgeable of online risks on Facebook.

From the users' perspective there may or may not be much difference between trusting the platform, and trusting information encountered on the platforms. In a similar manner, users may or may not be very nuanced about the exact target of regulation, or about what exactly their self-defense abilities are directed at. To address this uncertainty, we run all our models with three different dependent trust variables: trust in platforms, users, and information.

To account for the often-substantial variation in societal level trust we ran the survey in seven European countries: The Netherlands, Germany, France, Portugal, Hungary, Greece, and Estonia. We also needed to find a techno-social trust producing technology with comparable levels of penetration in these markets. Our choice was survey trust vis-à-vis social media, more concretely three Meta properties: Facebook, Instagram, and WhatsApp. Meta services are likely to be used in all the surveyed countries, unlike other platforms like Uber, AirBnB, or e-commerce marketplaces. Given Facebook's prominence and widely publicized scandals we also expected deeper familiarity platform related risks among our respondents. Last but not least, these services are the prime targets of a widely publicized European effort to bring very large digital online platforms under strong regulation.

Description of data

In collaboration with IPSOS Europe, we conducted the survey in The Netherlands, Portugal, Hungary, France, Estonia, Greece, and Germany between the 25th of February and 10th of March 2022. Our respondent panel consists of 1.000 completed online interviews from each country and is representative of the 18+ population with respect to age, education, and income. Our respondent selection criteria were that users have had a Facebook account in the last 12 months.

AS seen in Table 1. there substantial difference in the penetration of different social media platforms in our sample. The regional differences also visible for the three Meta properties: Facebook, Instagram ad WhatsApp.

Country	NL	PT	HU	FR	EE	GR	DE
Facebook	74%	84%	93%	68%	76%	86%	57%
WhatsApp	91%	86%	15%	44%	25%	30%	84%
Instagram	44%	67%	40%	34%	30%	68%	36%
Reddit	3%	6%	3%	2%	6%	7%	3%
Twitter	14%	20%	13%	16%	6%	30%	14%
TikTok	10%	25%	21%	13%	9%	36%	12%
YouTube	48%	65%	74%	44%	63%	83%	54%
Twitch	2%	3%	4%	4%	3%	8%	6%
Snapchat	11%	7%	9%	18%	6%	10%	8%
LinkedIn	27%	21%	8%	11%	8%	13%	10%
Telegram	6%	18%	5%	4%	5%	11%	14%
Signal	4%	1%	2%	3%	3%	3%	8%
None	3%	1%	2%	14%	10%	1%	5%

Table 1: Platform penetration by country

Given that the churn of younger Facebook users is well documented (Heath 2021), our respondents tend to be somewhat older, Greek users being the youngest (mean age=39.2, median age = 40), and Dutch users being the oldest (mean=49,5 median=52).

Trust in Meta platforms

We measured (both directly and indirectly) three trust variables which we'll use as dependent variables in our models. First, in line with the *trust in technology* approach, we measured users' trust in the Facebook platform. We also measured what kinds of risks the users associate with the platform.

Also, in line with the *trust by technology* approach, we measured to what extent users trust users and information they encounter on the platform. Besides direct measurement, we also asked multiple questions with regards to different aspects of potential risks, associated with these entities, such as manipulative user behavior, or information being harmful.

Direct trust measurements

The direct trust measurement shows substantial variation between the different trust objects, as well as the different platforms, and among the different countries. Facebook is the least trusted platform, and the least secure trust environment. On the other end, the messaging platform WhatsApp seems to be the most trusted, both as a platform, and as a place where information and users are trusted. Instagram lays somewhere in between.

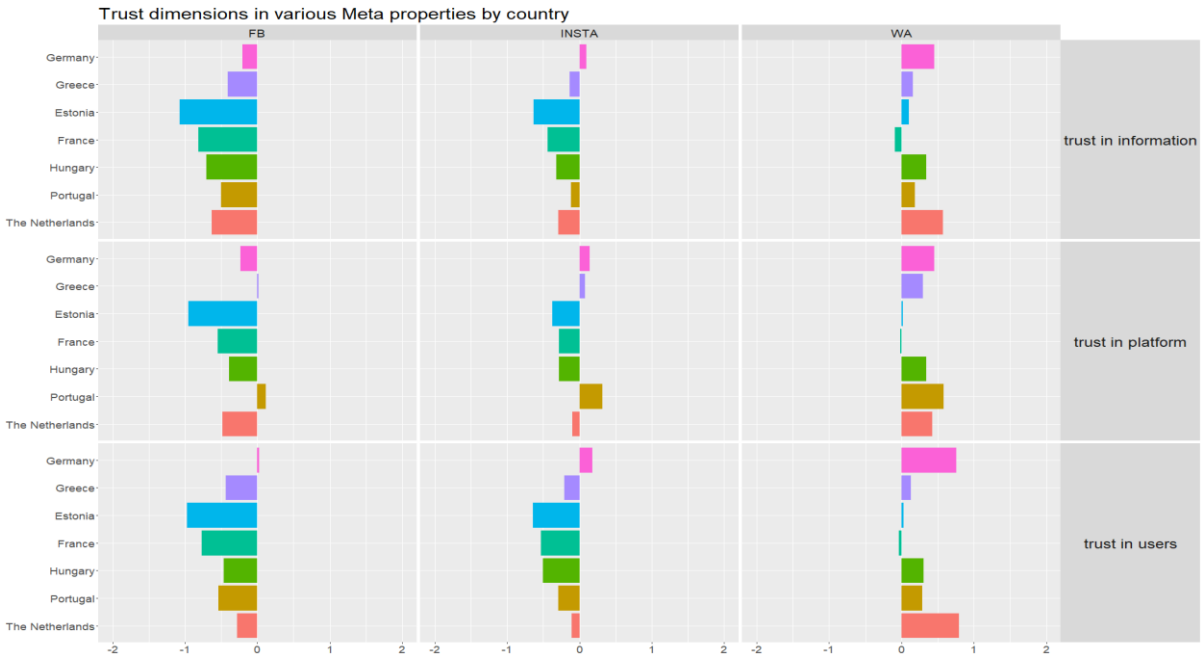

Figure 1: Trust on platforms, graph centered on the midpoint of a 7 point Likert scale.

In terms of country level differences, Estonians see Facebook the least trustworthy, Portuguese respondents see it the most trustworthy. The correlation coefficients between the perceived trustworthiness of the platform, the users, and the information suggest that these variables are sufficiently independent from each other to warrant being modelled individually.

service	platform_users	platform_information	users_information
Facebook	0.52	0.62	0.70
Instagram	0.64	0.70	0.76
WhatsApp	0.61	0.68	0.73

Table 2: correlation of various trust dimensions in different Meta platforms

For Facebook (and Instagram and Whatsapp, but not reported here) we also asked detailed questions about different factors related to trust and trustworthiness. We found that the detailed questions correlate very well with the one-shot trust questions.

Risk perceptions

Since trust is seen in the literature as a central instrument to engage with another in face of possible risks and harms, we extensively canvassed risk perceptions vis-à-vis both the platform, and the users and information encountered on the platform.

.

	NL	PT	HU	FR	EE	GR	DE
generic concerns							
I feel obliged to use FB	-1.07	-0.82	-0.76	-0.82	-1.48	-0.89	-0.81
I spend too much time on FB	-0.26	0.02	0.07	-0.36	-0.17	0.50	-0.71
others spend too much time on FB	1.11	1.61	1.55	1.16	1.58	1.63	0.95
FB is a safe environment to be in	-0.33	-0.14	-0.67	-0.72	-0.70	-0.37	-0.36
FB undermines relationships	0.05	0.46	0.53	0.37	-0.35	0.33	-0.02
Information flow related concerns							
concerned: filter bubbles	0.40	0.51	0.22	0.27	-0.17	0.70	0.03
concerned: harmful content	0.85	0.81	0.77	0.75	0.33	0.66	0.57
concerned: political censorship	0.31	0.27	0.65	0.33	-0.33	0.65	0.22
I can express myself on FB without fear	-0.13	-0.01	-0.39	-0.69	-0.64	-0.21	-0.37
Privacy related concerns							
concerned: generic privacy	0.68	0.75	0.59	0.69	0.31	0.85	0.63
concerned: privacy vs my government	-0.48	-0.08	0.27	0.02	-0.72	0.25	-0.49
concerned: privacy vs foreign gov	-0.26	0.16	0.11	0.09	-0.65	0.18	-0.32
concerned: privacy vs business	0.14	0.31	0.26	0.35	-0.28	0.31	-0.09
concerned: FB and gov cooperation	-0.28	0.32	0.50	0.18	-0.59	0.33	-0.36
Manupulation related concerns							
concerned: FB users misrepresent themselves	0.88	1.21	1.47	1.28	0.96	1.43	1.07
concerned: FB manipulation	-0.21	-0.06	-0.01	0.03	-0.68	0.00	-0.27
concerned: FB users manipulation	-0.50	-0.24	-0.22	-0.28	-0.86	0.17	-0.38
concerned: my government manipulation	-0.72	-0.30	0.19	-0.17	-1.37	0.02	-0.70
concerned: foreign gov manipulation	-0.33	0.05	-0.07	0.17	-1.01	0.25	-0.32
concerned: business manipulation	0.23	0.24	0.18	0.43	-0.59	0.35	-0.29

Table 3: Overuse, manipulation, information control, and privacy related risk perceptions on Facebook, by country. Means centered on the midpoint of 7p point Likert scale

We surveyed three different kinds of risk: those related to information flows, privacy, and manipulation. With regards to information flows, filter bubbles seem to be a shared concern in all countries except Estonia. Respondents are both worried about the spread of harmful content and political censorship, while they feel they cannot fully express themselves without fear on the platform. The political censorship concerns are strongest in Hungary and Greece, two countries with tumultuous internal politics, and lower general levels of trust in public institutions. In these countries concerns about local government surveillance is also noteworthy.

Privacy as a generic concern is widely shared in all countries. Interestingly, privacy concerns vis-à-vis concrete stakeholders do not correspond well to the generic concerns. In Estonia or Germany, the generic privacy concerns are high, but respondents were not that much concerned about government or business surveillance on the platform.

With regards to manipulation related concerns, there is a widely shared concern that users on FB misrepresent themselves, though this is usually not coupled with concerns about user manipulation. There are more concerns about manipulation by businesses, and in the case of Greece, France and Hungary, manipulation by foreign government. The authoritarian nature of the Hungarian government might explain why some Hungarian users are concerned about being manipulated by their own government. Also, the lack of concerns of Estonian respondents about possible Russian manipulation is noteworthy.

To contrast and contextualize these self-reported risk factors we also asked respondents if they were aware of any recent scandals involving Facebook. The answers suggest that only 10-25% of the respondents seem to have a clear recollection of one FB related scandal or another. In addition, there

seems to be a rather weak connection between risk perceptions and actual knowledge of FB scandals with Spearman's correlations ranging from the low of 0.03 between scandal knowledge and concerns about FB users misrepresent themselves, to the high of 0.19 between knowledge and FB users being manipulative.

In general, users don't seem to feel obliged to use Facebook because of some perceived social cost of exclusion (Tongia and Wilson 2011). This, coupled with the general feeling that FB is not a safe space, and that users (of course others) spend too much time there, suggests an environment where the stakes are not very high: respondents don't seem to feel trapped, and they may feel easy to leave, or reduce their engagement if and when they feel uncomfortable.

Trust pillar descriptives

Trust in self

We measured two aspects of self-trust: the self-assessed ability to recognize risks, and the perceived ability to protect oneself against those risks.

	NL	PT	HU	FR	EE	GR	DE
Risk recognition related variables							
recognise: manipulation by FB	-0.28	0.50	0.74	0.12	-0.15	0.83	0.37
recognise: manipulation by national gov	-0.48	0.20	0.73	-0.09	-0.33	0.57	0.06
recognise: manipulation by other user	-0.08	0.74	0.90	0.29	0.08	0.96	0.55
recognise: trustworthy info	-0.07	0.72	0.80	0.31	0.29	0.89	0.54
recognise: trustworthy user	-0.07	0.62	0.78	0.28	0.30	0.70	0.56
Risk protection related variables							
protect myself: being abused	-0.17	0.13	0.23	-0.13	-0.05	0.40	0.09
protect myself: censorship	-0.32	0.02	0.18	-0.08	-0.50	0.33	-0.01
protect myself: filter bubbles	-0.32	0.07	0.45	-0.12	-0.32	0.39	0.18
protect myself: harmful content	0.06	0.37	0.75	0.15	0.05	0.74	0.38
protect myself: from surveillance	-0.68	-0.29	-0.09	-0.34	0.00	0.10	-0.13

Table 4: self-trust dimensions by country (Means centered on 7 point Likert scale midpoint)

While the mean responses to the risk recognition questions seem to be slightly above the midpoint for most countries, self-protection responses were more varied, with Estonian, French, and Dutch respondents the least confident about being able to protect themselves from manipulation, filter bubbles and censorship.

For the purposes of the analysis, we created two composite variables from these questions, one for threat recognition, another for the ability to self-protect. The internal consistency of these variables was adequate (Cronbach's alphas 0.86, and 0.84, respectively). The two composite variable's correlation is 0.68.

To contrast these responses with actual, self-reported protection measures, we asked respondents about how often users actually take self-defense action on the platform, such as restricting the visibility of their own posts, knowing about, and using content reporting tools on the platform. Though ~64% of respondents have heard about content reporting tools, (with Portugal (77%) and France (45%) at the two extremes), only 38% of all the respondents use harmful content reporting tools more often than not, and only 14,5% use them every time they see content which they deem harmful.

Trust in platform

We measured expectations vis-à-vis the trust in the platform's self-regulating capacity with several different questions over the platform's perceived ability, benevolence, and integrity (Mayer et al. 1995).

The general trustworthiness assessment of the Facebook platform is rather negative, with response means well below the midpoint. This is especially true for procedural fairness indicators ("fair", "transparent"), as well as benevolence indicators ("acts in users' best interest", "makes the world a better place", "removes fraud users" etc.), while the variables on Facebook's competence and the quality of its services perform slightly better.

	NL	PT	HU	FR	EE	GR	DE
Generic trustworthiness factors							
fair	-0.63	-0.08	-0.30	-0.56	-0.91	-0.32	-0.18
transparent	-0.85	-0.34	-0.09	-0.82	-1.17	-0.42	-0.39
acts in user best Interest	-1.00	-0.23	-0.55	-0.79	-1.13	-0.61	-0.44
provides high quality services	-0.20	0.22	0.15	-0.20	-0.42	0.11	-0.09
improved safety reliability from the past	-0.24	0.08	0.16	-0.35	-0.61	-0.10	-0.11
makes the world a better place	-0.95	-0.24	-0.31	-0.77	-1.04	-0.29	-0.52
over controlling, over regulating	-0.07	0.49	0.26	-0.30	-0.31	0.40	-0.01
User related trustworthiness factors							
removes fraud users	-0.01	0.23	-0.08	-0.52	-0.89	-0.17	-0.03
prevents user manipulation	-0.33	0.01	-0.32	-0.64	-1.20	-0.46	-0.26
prevents government manipulation	-0.29	-0.12	-0.38	-0.70	-1.18	-0.68	-0.24
Information related trustorthiness factors							
trustworthy free speech space	-0.57	0.08	-0.08	-0.54	-0.94	-0.17	-0.33
trustworthy friend info source	-0.34	0.08	0.00	-0.23	-0.37	0.03	0.03
trustworthy pol info source	-1.15	-0.52	-0.84	-0.85	-1.46	-0.61	-0.59
helps identify bad content	-0.06	0.31	-0.23	-0.43	-0.97	-0.16	-0.04
Privacy related trustworthiness factor							
trustworthy personal data fiduciary	-0.57	-0.55	-0.38	-0.83	-1.16	-0.49	-0.32
prevents government spying	-0.32	-0.18	-0.40	-0.79	-1.29	-0.78	-0.29

Table 5: Trust in FB self-regulating capacity (means centered on 7 point Likert scale midpoint)

Respondents demonstrated high levels of skepticism in terms of whether FB can be trusted to create a space by addressing user, information, and privacy related concerns.

Again, to contrast these responses with actual knowledge, we asked users whether they have heard about, and interacted with the Facebook Oversight Board, an important institution of FB self-governance. Depending on the country, at least 75% of our respondents never heard of the FB oversight board, don't know what it does, or how to use it.

Institutional pillar

In recent years the political pressure to regulate large, mostly US-based online platforms has been growing, both at the national as well as at the EU level. Platforms' Terms of Service, describing the rules on the platform do not always reflect the laws of the land in various jurisdictions (Suzor 2019).

As a result, platforms both may allow content which would be illegal in that particular jurisdiction, and filter content which would fall under protected speech. It is often unclear to what extent the problems experienced in content moderation are the result of deliberate decisions by platform operators, as opposed to the inherent limitations of automated content moderation, and the high costs of human overview. The opacity of content moderation makes it difficult for both individual users, as well as for the society to fully grasp how such systems are shaping information flows (Bodó et al. 2017; Gillespie 2018).

In response, a growing body of national and EU courts and legislators address these issues (De Streel et al. 2020). Laws in France, the UK and Germany spelled out obligations which hoped to make platforms a more trustworthy actor in the digital society. Most recently the European Digital Services Act defines wide ranging rules to combat risks and harms related to the spread of misinformation, propaganda, illegal, or simply harmful content, as well as creating increased transparency in how content moderation takes place.

In the light of these regulatory efforts, it is crucial to understand to what extent platform users trust the government (national and EU) to make sure that the platforms are trustworthy.

Country	NL	PT	HU	FR	EE	GR	DE
trust in police	2.60	1.64	1.23	1.84	3.15	1.30	2.47
trust in politicians	0.66	-0.14	-0.80	-0.51	0.14	-0.90	0.46
trust in government	1.29	0.69	-0.27	0.18	1.26	-0.43	1.09
trust in legal system	2.02	1.05	0.40	0.86	2.48	0.73	1.81
trust in regulating FB							
EU has the benevolence	0.16	-0.20	-0.67	-0.03	0.05	-1.00	0.00
EU has the strength	0.20	0.53	0.21	0.33	-0.15	0.53	0.41
national government has the benevolence	0.20	-0.45	-1.35	0.04	0.11	-1.26	0.05
national government has the strength	-0.12	-0.21	-0.44	0.27	-1.11	-0.15	0.23

Table 6: Trust in governments' platform regulation powers

We broke down this question into two sets of variables. On the one hand, general institutional trust measures generic trust versus various local public institutions. On the other hand, a set of questions measures to what extent citizens think the local or European government has the strength and the benevolence to regulate Facebook. We define benevolence as the regulator acting in the public interest as opposed to some particular interest. Interestingly in those countries which have enacted platform specific regulation (Germany: NetzDG, France- AVIA) respondents saw their governments being strong enough to regulate, though there is no strong sign about the positive impact of such regulation. On the other hand, Hungarian and Greek respondents were especially distrustful that their government would regulate in the public interest. Expectations with regards to the strength and benevolence of EU institutions are higher.

Explanatory analysis

We test our hypotheses by OLS regression models using countries as fixed-effect predictors. Three models are fitted to each dependent variable (trust in FB as a platform, trust in users, trust in the information on the platform). The first model includes only direct effects. The second model adds pillar-related, and the third adds risk-related interaction terms. In line with our hypotheses, we

included as explanatory variables the indicators on perceived risks, users' self-confidence in recognizing and countering risks, the perception on FB's self-regulatory efforts and attitudes about the government's responsibility on regulating the platform. The usual socio-demographic indicators, variables on different kinds of trust as well as on political interest and ideological orientation were included as controls. Our findings are shown in Table 6.

	Facebook as a platform			Facebook users			Facebook information		
Predictors	std. Beta	std. Beta	std. Beta	std. Beta	std. Beta	std. Beta	std. Beta	std. Beta	std. Beta
(Intercept)	-0.17 *** (-0.23 – -0.11)	-0.18 *** (-0.24 – -0.13)	-0.16 *** (-0.22 – -0.11)	0.14 *** (0.07 – 0.20)	0.12 *** (0.06 – 0.18)	0.14 *** (0.07 – 0.20)	-0.04 (-0.11 – 0.02)	-0.06 (-0.12 – 0.00)	-0.04 (-0.10 – 0.02)
risk-perception	-0.13 *** (-0.15 – -0.11)	-0.13 *** (-0.16 – -0.11)	-0.13 *** (-0.15 – -0.11)	-0.02 (-0.04 – 0.01)	-0.02 (-0.04 – 0.00)	-0.02 (-0.04 – 0.01)	-0.02 (-0.04 – 0.00)	-0.02 * (-0.04 – -0.00)	-0.02 (-0.04 – 0.00)
agency: recognition	0.01 (-0.01 – 0.04)	0 (-0.02 – 0.03)	0.02 (-0.01 – 0.04)	0.05 *** (0.02 – 0.08)	0.03 (-0.00 – 0.06)	0.05 *** (0.02 – 0.08)	-0.02 (-0.05 – 0.01)	-0.03 (-0.06 – 0.00)	-0.02 (-0.05 – 0.01)
agency: protection	0.14 *** (0.11 – 0.16)	0.14 *** (0.12 – 0.17)	0.13 *** (0.10 – 0.16)	0.08 *** (0.05 – 0.11)	0.11 *** (0.08 – 0.14)	0.08 *** (0.05 – 0.11)	0.16 *** (0.13 – 0.19)	0.17 *** (0.14 – 0.20)	0.16 *** (0.13 – 0.19)
FB self-regulation	0.43 *** (0.40 – 0.45)	0.43 *** (0.40 – 0.45)	0.42 *** (0.40 – 0.45)	0.35 *** (0.33 – 0.38)	0.35 *** (0.33 – 0.38)	0.35 *** (0.33 – 0.38)	0.41 *** (0.38 – 0.43)	0.40 *** (0.38 – 0.43)	0.41 *** (0.38 – 0.43)
government responsibility	-0.01 (-0.03 – 0.01)	-0.01 (-0.03 – 0.01)	-0.01 (-0.03 – 0.01)	0 (-0.03 – 0.02)	0 (-0.03 – 0.02)	0 (-0.03 – 0.02)	-0.01 (-0.03 – 0.01)	-0.01 (-0.03 – 0.01)	-0.01 (-0.03 – 0.01)
FB scandal knowledge	-0.05 *** (-0.07 – -0.03)	-0.06 *** (-0.08 – -0.04)	-0.05 *** (-0.07 – -0.03)	-0.02 * (-0.05 – -0.00)	-0.03 * (-0.05 – -0.01)	-0.02 * (-0.05 – -0.00)	-0.02 (-0.04 – 0.01)	-0.02 (-0.04 – 0.00)	-0.02 (-0.04 – 0.01)
institutional trust	0.06 *** (0.03 – 0.08)	0.06 *** (0.03 – 0.08)	0.06 *** (0.03 – 0.08)	0.04 ** (0.01 – 0.07)	0.04 ** (0.01 – 0.07)	0.04 ** (0.01 – 0.07)	0.04 ** (0.02 – 0.07)	0.04 ** (0.01 – 0.06)	0.04 ** (0.01 – 0.07)
technology trust	0.14 *** (0.12 – 0.17)	0.14 *** (0.12 – 0.16)	0.14 *** (0.12 – 0.17)	0.04 *** (0.02 – 0.07)	0.04 *** (0.02 – 0.06)	0.04 *** (0.02 – 0.07)	0.06 *** (0.03 – 0.08)	0.06 *** (0.03 – 0.08)	0.06 *** (0.03 – 0.08)
generalized trust	0.05 *** (0.03 – 0.07)	0.05 *** (0.03 – 0.08)	0.05 *** (0.03 – 0.07)	0.22 *** (0.20 – 0.25)	0.23 *** (0.20 – 0.25)	0.23 *** (0.20 – 0.25)	0.15 *** (0.13 – 0.18)	0.16 *** (0.13 – 0.18)	0.15 *** (0.13 – 0.18)
ideology (+right)	-0.01 (-0.03 – 0.01)	-0.02 (-0.04 – 0.00)	-0.01 (-0.03 – 0.01)	0.01 (-0.01 – 0.03)	0.01 (-0.01 – 0.03)	0.01 (-0.01 – 0.03)	0 (-0.02 – 0.02)	0 (-0.02 – 0.02)	0 (-0.02 – 0.03)
political interest	-0.01 (-0.03 – 0.02)	-0.01 (-0.03 – 0.02)	0 (-0.03 – 0.02)	-0.04 ** (-0.06 – -0.01)	-0.04 ** (-0.06 – -0.02)	-0.04 ** (-0.06 – -0.01)	-0.03 ** (-0.05 – -0.01)	-0.03 * (-0.05 – -0.01)	-0.03 * (-0.05 – -0.01)
age	0.02 (-0.00 – 0.04)	0.02 (-0.00 – 0.04)	0.02 (-0.00 – 0.04)	0.01 (-0.01 – 0.03)	0.01 (-0.01 – 0.03)	0.01 (-0.01 – 0.03)	0 (-0.02 – 0.02)	0 (-0.02 – 0.02)	0 (-0.02 – 0.02)
gender (1 - female)	0.02 (-0.02 – 0.06)	0.02 (-0.02 – 0.06)	0.02 (-0.02 – 0.06)	-0.09 *** (-0.13 – -0.04)	-0.08 *** (-0.13 – -0.04)	-0.09 *** (-0.13 – -0.04)	-0.03 (-0.07 – 0.01)	-0.03 (-0.07 – 0.01)	-0.03 (-0.07 – 0.01)
education	-0.03 * (-0.05 – -0.01)	-0.02 * (-0.04 – -0.00)	-0.03 * (-0.05 – -0.01)	-0.04 *** (-0.06 – -0.02)	-0.03 ** (-0.06 – -0.01)	-0.04 *** (-0.06 – -0.02)	-0.04 ** (-0.06 – -0.01)	-0.03 ** (-0.06 – -0.01)	-0.04 ** (-0.06 – -0.01)
income	-0.02 (-0.04 – 0.00)	-0.01 (-0.03 – 0.01)	-0.02 (-0.04 – 0.00)	-0.02 (-0.04 – 0.00)	-0.02 (-0.04 – 0.01)	-0.02 (-0.04 – 0.00)	-0.02 * (-0.05 – -0.00)	-0.02 * (-0.04 – -0.00)	-0.02 * (-0.05 – -0.00)
domicil	0 (-0.02 – 0.02)	0 (-0.02 – 0.03)	0.01 (-0.02 – 0.03)	0.01 (-0.01 – 0.03)	0.01 (-0.01 – 0.03)	0.01 (-0.01 – 0.03)	0.02 (-0.01 – 0.04)	0.02 (-0.00 – 0.04)	0.02 (-0.01 – 0.04)
agency: recognition * FB self-regulation		-0.01 (-0.03 – 0.02)			-0.07 *** (-0.10 – -0.05)			-0.01 (-0.04 – 0.01)	
agency: recognition * government responsibility		-0.02 (-0.04 – 0.01)			-0.01 (-0.04 – 0.01)			-0.02 (-0.04 – 0.00)	
agency: protection * FB self-regulation		0.04 *** (0.02 – 0.07)			0.09 *** (0.07 – 0.12)			0.05 *** (0.02 – 0.07)	
agency: protection * government responsibility		0 (-0.02 – 0.02)			0.03 * (0.01 – 0.06)			0.06 *** (0.03 – 0.08)	
FB self-regulation * government responsibility		0.03 ** (0.01 – 0.04)			0 (-0.02 – 0.02)			-0.01 (-0.02 – 0.01)	
risk-perception * agency: recognition			-0.05 *** (-0.07 – -0.02)			0 (-0.02 – 0.03)			-0.02 (-0.04 – 0.01)
risk-perception * agency: protection			0.03 ** (0.01 – 0.06)			0 (-0.03 – 0.02)			0.02 (-0.00 – 0.05)
risk-perception * FB self-regulation			0.01 (-0.01 – 0.03)			-0.01 (-0.03 – 0.01)			0.01 (-0.01 – 0.03)
risk-perception * government responsibility			0 (-0.01 – 0.02)			-0.01 (-0.03 – 0.01)			0 (-0.02 – 0.02)
Observations	6252	6252	6252	6252	6252	6252	6252	6252	6252
R2 / R2 adjusted	0.418 / 0.416	0.422 / 0.419	0.420 / 0.418	0.313 / 0.311	0.320 / 0.318	0.314 / 0.311	0.350 / 0.348	0.356 / 0.353	0.351 / 0.348

*p<0.05 **p<0.01 ***p<0.001

Table 7: OLS regression models

Interestingly, the risk perceived in relation with Facebook use decreases only the trust in the platform, but not in its users or information (H1 is partially supported). This fact indicates that those who feel that Facebook use involves several risks blame the platform for these, but still have trust in the content it mediates. It also points to the fact that risks perceptions remain at a generalized level and are not related to concrete potential issues with users or with information. It seems that a higher level of awareness does not make people more skeptical or suspicious of what they see on the platform even if they have a less favorable view about the platform as a whole. One reason for that can be that they rely on one of the pillars to handle those risks.

Our findings also suggest that if people think that they can be protected from harm, they have a higher level of trust in Facebook. Protection can come from different sources, but they are not equally effective in building confidence. Trust in Facebook's self-regulation capacity is clearly the most important factor: if people believe that Facebook will protect them from harm, they trust more in the platform, its users, and the information (H3 is supported). Feeling of agency also uniformly affects all dimensions of Facebook-related trust: those who are confident that they can protect themselves from the platform related risks trust the platform, its users and information (H2b) more. However, the effect size is weaker than the effect of the trust in Facebook's self-regulation capacity. All else being equal, there are approximately 2.2 - 2.6 points difference in the different trust dimensions between those who extremely believe in and those who highly skeptical of Facebook's capacity and willingness to protect users. In case of self-protection, the maximum difference is 0.6 - 1.1 points (see Figure 2).

It is also important to note that agency must be active to build trust effectively. The ability to recognize harms can only weakly increase trust in users (H2a is supported in case of users), but it is not significantly associated with trust in the platform and its information (H2a is rejected in case of platform and information). This implies that recognition per se is not protection, the awareness of harms is not sufficient to develop trust toward a platform.

Nonetheless, our evidence also points out that trust cannot be built from 'outside' the platform: people do not base their Facebook trust on external institutional actors' regulatory capacities. The level of trust the government's regulatory capacity is not associated with people' trust in Facebook (H4 is rejected). It seems that government is not the external actor which could help to build platform-related trust. This finding can be related to our observation discussed above that people are not very convinced about the benevolence of the regulatory efforts of their governments.

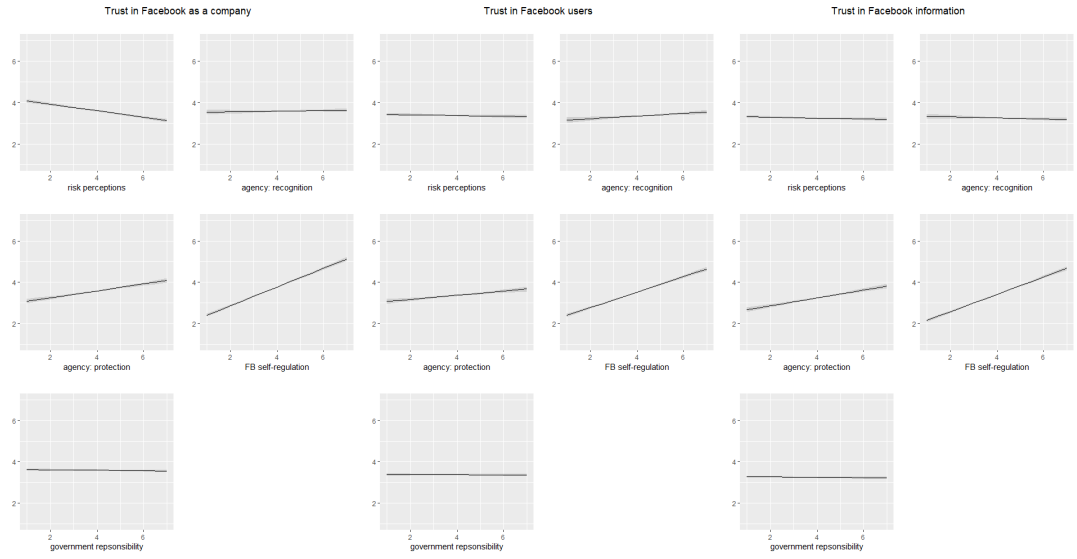

Figure 2: Main effects of our independent variables in the regression models.

When it comes to the interaction effects, we find interesting patterns. There is indeed a moderated relationship between the two key predictors of trust, namely belief in Facebook's self-regulatory capacity and agency of self-protection, but its direction is the opposite of what we expected (Figure 3). Surprisingly, there is an amplification effect rather than a substitution one. The effect of agency is higher when people believe that Facebook also protects them. In other words, people trust in Facebook more when they are confident that Facebook does its best to protect them, and they are also able to protect themselves. In this case the two protection mechanisms are connected and help to build trust on the platform. In contrast, self-confidence matters less when the platform is experienced to do little to keep its users safe. Even if people think they can protect themselves, they cannot trust in the platform absent of such efforts. This amplifying effect is prevalent in each dimension of Facebook-trust. This fact also shows that the perceptions about Facebook's self-regulatory capacity is the key in understanding platform-related trust.

The perceived level of government responsibility is also significantly interacted with the two key independent variables in an amplifying way, but these effects are not uniform across the dimensions and also much weaker. Government's responsibility can slightly contribute to trust in the platform when people perceive the platform to make effective protective measures, and self-protective agency also have stronger effect on trust in users and information when the perceived level of government responsibility is higher.

We found, however, one interaction effect which supports our original hypothesis. As we discussed above, the ability to recognize harms does not contribute to trust in the platform. However, when people don't feel that Facebook protects them, the role of this ability in trusting users is enhanced. In this case, agency can slightly substantiate the positive effect of Facebook's self-regulatory efforts: when people know that Facebook does not help them to identify threats, their confidence in their own abilities to recognize these can contribute to higher trust in users (H5 is rejected in all but one cases).

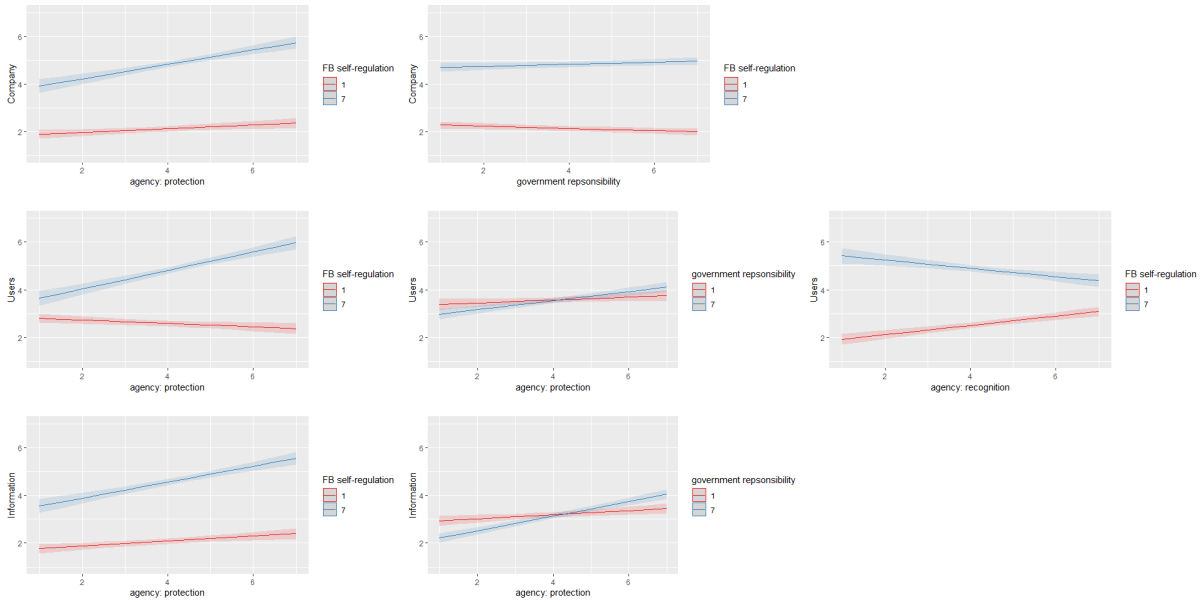

Figure 3: Significant interaction effects of the pillars of trust in the regression models.

Last, we tried to find evidence that these mechanisms work differently for people who are aware of the risks associated with Facebook use (Figure 4). We argued that the belief in the protective capacities of different agencies has different meanings for people who are aware of the risks compared to those

who are not. Our findings show that the differences are minor, most effects we found are uniform across people with different risk perceptions. The only exceptions are related to the two agency variables in case of platform trust. However, the directions of these interaction effects are different. It seems that for people with higher risk-awareness, the perceived ability to identify the threats even decreases the trust in the platform, while people who feel Facebook to be a less risky place can build trust from their confidence in recognizing harms. In other words, if there is nothing to worry about, trust can be built from the perceived ability to be vigilant. On the other hand, for people who think that Facebook is a dangerous place, recognition of the threats is not a comfortable experience, instead it is something that makes people more skeptical about the platform. The ability of self-protection, however, seems to be more important in developing platform-related trust when Facebook is perceived as risky. It is easy to see that if there are no threats at all, self-protection ability does not seem to be a very useful skill. However, it is more appreciated when risks are clearly perceived. This case shows that the reflective self-protection ability is different from the unreflective one, but also demonstrates that recognition and protection are separate dimensions of personal agency with distinct effects on platform-related trust.

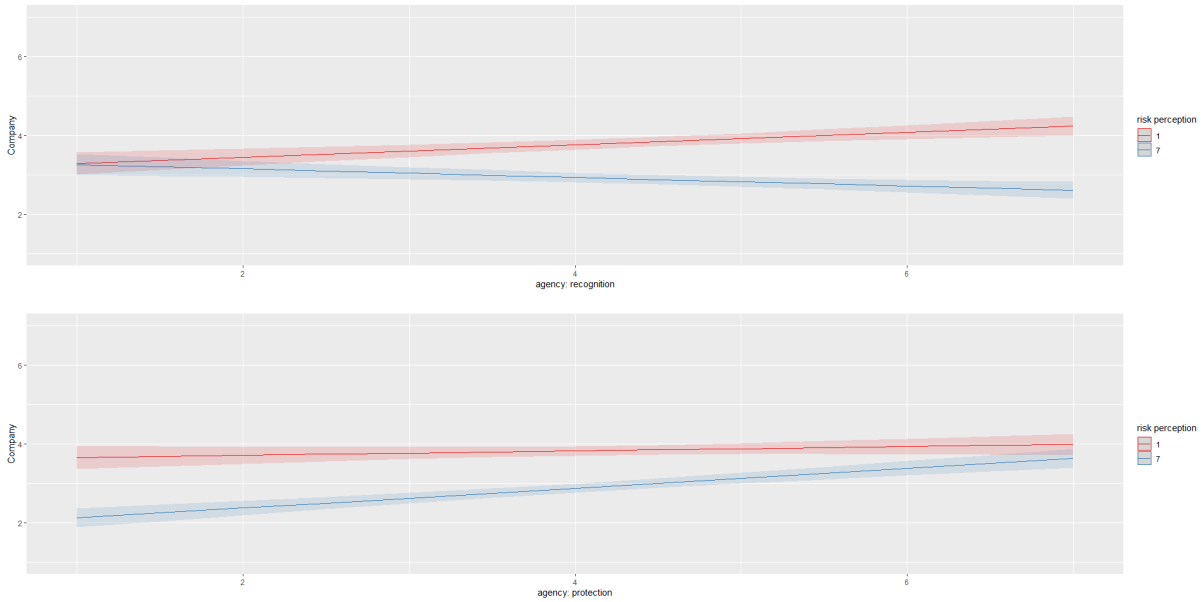

Figure 4: Significant interaction effects of the pillars of trust and risk perceptions in the regression models.

Discussion

Our findings, both the descriptive data and the explanatory models paint a complex, multi-dimensional picture around the (political) economy of trust in our digital societies, where various private, technological intermediaries, such as Meta platforms play a central role in trust interpersonal relations. We have identified several issues relevant for individual users, platform operators, as well as policymakers.

First, platforms, in our case Facebook are seen by our respondents in each country as entities which introduce several individual and societal level risks and harms into social relations. On the one hand,

this recognition is the first step in the right direction: individuals use these technologies aware of the risks, and societies can start looking for solutions to minimize harms. On the other hand, despite the fact that Facebook has been mired in numerous scandals, such as the Cambridge Analytica case, knowledge does not translate directly into risk perceptions, raising the question on what these perceptions are really based on: first-hand experience, or on some more generic, imprecise, folk theories of how this environment operates.

The generic risk perceptions, and the users' self-confidence to recognize threats also play a complex role in users' trust in and the platform. Trust in the platform decreases if users associate more risk with it, or when risk aware users feel confident in recognizing the threats; and grows if they feel confident that they can protect themselves. But risk only plays a role in the trust *in* the platform and not *on* the platform, despite the fact that the sources of actual risks are other users, or contentious information. This may mean that risk assessment and recognition take place at the generic level, and this platform level risk then isn't necessarily qualified further on a case-by-case basis vis-à-vis individual users and pieces of information.

In fact, the largest effect sizes were associated with our generic trust measures: generalized trust in the case of trust on the platform in users and information, and technology trust for the trust in platform. This also means that trust *in* the platform seems to be more strongly defined by these generic trust attitudes, than concrete, platform-specific risk assessments. This may easily be an artifact of the past, where the public discourse around the societal impact of digital technologies was rather enthusiastic, fueled an uncritical acceptance of the techno-solutionist agenda (Morozov 2014). From a normative perspective, it seems more desirable that trust in technology is based on a reflexive, rational understanding of the pros and cons of a techno-social system, than a blind faith in its promised, but never-quite-realized positive impact, and a blind eye towards the harms it may cause.

Second, out of the three pillars of trustworthiness the platforms' self-regulation is the strongest, followed by users' confidence in being able to protect themselves. These findings, for us, are massive red flags, which point to potentially dangerous misplaced trust in technology. Users may rightfully expect tech companies to act in their (i.e., their users') best interest, and commit sufficient resources to make these platforms a safe space. It is also clear that there is little empirical evidence that platforms are actually benevolent, and are able or willing to act with what some call fiduciary care towards their users (Balkin 2016; Bodó 2020). There is also ample evidence that platforms have repeatedly and gravely breached the trust of their users, by actively acting against their interest. Users trust a provably untrustworthy agent to act in their best interest. This is the textbook case of potentially extremely harmful misplaced or unreflective trust.

In face of an untrustworthy actor, self-confidence in one's ability to protect themselves is a key resource. Without it, one is at the mercy of the elements, with it one may stand a chance to be the master of their own destiny. The significant and relatively strong effects of one's ability to protect themselves may well be read as a good sign. Our optimism, however, is heavily curtailed by the very limited traces of actual self-protecting actions. When asked, users hardly say they take active steps to manage their information environments using the tools given to them, and rarely they give signals to the platform to do so. These findings, taken together, raise the prospect of a truly frightening scenario, in which users deluding themselves of being in control, when in fact they aren't.

Our data suggests that actually both of the pillars on which they base their trust in the platform may be delusional: they trust the platform to protect them, which it won't, and they think they can protect themselves, which they don't. They may feel safe, but in reality, they are standing naked in the storm.

Third, the current institutional order positions the EU as the sole actor which both has the power and the expertise to reign in global digital service providers, and force them to act in an ethical, law compliant, transparent, accountable, and trustworthy manner. The almost complete lack of trust in the EU to provide effective trustworthiness safeguards versus platforms is somewhat disturbing. The European Commission has spent enormous amounts of time, expertise, and political capital to pass a comprehensive set of regulations that try to regulate online platforms. Apparently, these efforts have not been visible, legible, or relevant for our respondents. Policymakers should take note of this finding.

Finally, the turmoil around Twitter after Elon Musk has taken it private provides us with an interesting counterfactual. The apparent mass exodus of users was prompted by the radical changes in perceived benevolence of the platform. The two other pillars: users' ability to recognize and address threats, and the regulatory environment hasn't changed, only the platform's expected behavior. Also, the changes in the ownership structure may have shifted and focused the risk perceptions and trusting attitudes from abstract, generic, and nebulous, to the concrete, specific, and relational. This also suggests that - at least in this domain – even if there are multiple frameworks on which trust may rest, the collapse of one is enough to destroy everything.

Limitations and further research

The study has several limitations. It has been a one-shot study in a rapidly environment. Also, we surveyed trust attitudes, something notoriously difficult to measure. These factors introduce potentially many unknown biases into our findings. As the fate of Twitter should warn us trust is hard to build but collapses extremely fast. Our findings are probably less about long term trends than the tensions within complex systems which may hold for quite a long time before undergoing rapid, cascading transformations.

Acknowledgments

The fieldwork was executed by the Ipsos Research Setup on the Ipsos panel, administered by Ipsos Interactive Services. We thank Dr. Theo Araujo for his help in developing the questionnaire. We would like to express our gratitude to Dr. Piret Ehin, who provided us valuable insight into to seemingly idiosyncratic data from Estonia. In this case her insight was that Estonian users may very well be aware of what to expect on Facebook in terms of Russian propaganda, meaning that the lack of concern is due to a certain certainty in terms of negative expectations.

Bodo received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 759681. Márton Bene is of the Bolyai János Research Fellowship awarded by the Hungarian Academy of Sciences (BO/334_20).

References

Anon. 2015. “The Trust Machine - The Promise of the Blockchain.” *The Economist*, October 31.

Balkin, Jack M. 2016. “Information Fiduciaries and the First Amendment.” *U.C. Davis Law Review* 49(4):1183.

Bargain, Olivier, and Ulugbek Aminjonov. 2020. “Trust and Compliance to Public Health Policies in Times of COVID-19.” *Journal of Public Economics* 192:104316. doi: 10.1016/j.jpubeco.2020.104316.

Bodó, Balázs. 2020. “Mediated Trust: A Theoretical Framework to Address the Trustworthiness of Technological Trust Mediators.” *New Media & Society* 146144482093992. doi: 10.1177/1461444820939922.

Bodó, Balázs. 2021. “The Commodification of Trust.” *SSRN Electronic Journal*. doi: 10.2139/ssrn.3843707.

Bodó, Balázs, Natali Helberger, Kristina Irion, Frederik J. Borgesius Zuiderveen, Judith Moller, Bob van der Velde, Nadine Bol, Bram van Es, and Claes H. de Vreese. 2017. “Tackling the Algorithmic Control Crisis – the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents.” *Yale Journal of Law & Technology* 19:133.

Bodó, Balazs, and Heleen Janssen. 2022. “Maintaining Trust in a Technologized Public Sector.” *Policy & Society* 41(3):414–29. doi: https://doi.org/10.1093/polsoc/puac019.

Bonilla-Zorita, Gabriel, Mark D. Griffiths, and Daria J. Kuss. 2020. “Online Dating and Problematic Use: A Systematic Review.” *International Journal of Mental Health and Addiction*. doi: 10.1007/s11469-020-00318-9.

De Streel, Alexandre, Elise Defreyne, Hervé Jacquemin, Michèle Ledger, Alejandra Michel, Alessandra Innesti, Marion Goubet, and Dawid Ustowski. 2020. *Online Platforms’ Moderation of Illegal Content Online*. Policy Department for Economic, Scientific and Quality of Life Policies Directorate-General for Internal Policies.

Fukuyama, Francis. 1995. *Trust: The Social Virtues and the Creation of Prosperity*. New York, NY: Free press.

Giddens, Anthony. 1990. *The Consequences of Modernity*. Cambridge, UK: Polity Press.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Hardin, Russell. 2002. *Trust and Trustworthiness*. New York: Russell Sage Foundation.

Hawdon, James. 2008. “Legitimacy, Trust, Social Capital, and Policing Styles: A Theoretical Statement.” *Police Quarterly* 11(2):182–201. doi: 10.1177/1098611107311852.

Heath, Alex. 2021. “Facebook’s Lost Generation.” *The Verge*. Retrieved December 20, 2022 (https://www.theverge.com/22743744/facebook-teen-usage-decline-frances-haugen-leaks).

Helberger, Natali. 2019. “On the Democratic Role of News Recommenders.” *Digital Journalism* 7(8):993–1012. doi: 10.1080/21670811.2019.1623700.

Kann, Sharon, and Angelo Carusone. 2022. “In Less than a Month, Elon Musk Has Driven Away Half of Twitter’s Top 100 Advertisers.” *Media Matters for America*. Retrieved February 1, 2023 (https://www.mediamatters.org/elon-musk/less-month-elon-musk-has-driven-away-half-twitters-top-100-advertisers).

Keller, Tobias R., and Ulrike Klinger. 2019. “Social Bots in Election Campaigns: Theoretical, Empirical, and Methodological Implications.” *Political Communication* 36(1):171–89. doi: 10.1080/10584609.2018.1526238.

Khan, Abdul Kader. 2022. “Trust in Artificial Intelligence: Toward Measuring the Impact of Public Perception.” *AMCIS 2022 TREOs*.

Lagerspetz, Olli. 1998. *Trust: The Tacit Demand*. Vol. 1. Dordrecht: Springer Netherlands.

Levi, Margaret, Audrey Sacks, and Tom Tyler. 2009. “Conceptualizing Legitimacy, Measuring Legitimating Beliefs.” *American Behavioral Scientist* 53(3):354–75. doi: 10.1177/0002764209338797.

Luhmann, Niklas. 2017. *Risk: A Sociological Theory*.

Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. “An Integrative Model of Organizational Trust.” *The Academy of Management Review* 20(3):709–34. doi: 10.2307/258792.

Morozov, Evgeny. 2014. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: PublicAffairs.

Murphy, Kristina. 2005. “Regulating More Effectively: The Relationship between Procedural Justice, Legitimacy, and Tax Non-Compliance.” *Journal of Law and Society* 32(4):562–89. doi: 10.1111/j.1467-6478.2005.00338.x.

Norris, Pippa. 2022. *In Praise of Skepticism: Trust but Verify*. New York, NY: Oxford University Press.

O’Neill, Onora. 2002. *Autonomy and Trust in Bioethics*. Cambridge ; New York: Cambridge University Press.

Ortega, Josue, and Philipp Hergovich. 2018. “The Strength of Absent Ties: Social Integration via Online Dating.” *ArXiv:1709.10478 [Physics, q-Fin]*.

Quintais, Joao Pedro, Balázs Bodó, Alexandra Giannopoulou, and Valeria Ferrari. 2019. “BLOCKCHAIN AND THE LAW: A CRITICAL EVALUATION (Book Review).” *Stanford Journal of Blockchain Law & Policy* 2(1).

Smith, Kevin B., Christopher W. Larimer, Levente Littvay, and John R. Hibbing. 2007. “Evolutionary Theory and Political Leadership: Why Certain People Do Not Trust Decision Makers.” *The Journal of Politics* 69(2):285–99. doi: 10.1111/j.1468-2508.2007.00532.x.

Spaiser, Viktoria, Thomas Chadefaux, Karsten Donnay, Fabian Russmann, and Dirk Helbing. 2017. "Communication Power Struggles on Social Media: A Case Study of the 2011–12 Russian Protests." *Journal of Information Technology & Politics* 14(2):132–53. doi: 10.1080/19331681.2017.1308288.

Stokel-Walker, Chris. 2022. "Twitter May Have Lost More than a Million Users since Elon Musk Took Over." *MIT Technology Review*. Retrieved February 1, 2023 (https://www.technologyreview.com/2022/11/03/1062752/twitter-may-have-lost-more-than-a-million-users-since-elon-musk-took-over/).

Suzor, Nicolas P. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge, UK: Cambridge University Press.

Sztompka, Piotr. 1999. *Trust: A Sociological Theory*. Cambridge, UK: Cambridge University Press.

Tongia, Rahul, and Ernest J. III Wilson. 2011. "The Flip Side of Metcalfe's Law: Multiple and Growing Costs of Network Exclusion." *International Journal of Communication* 5(0):17.

Wright, A., and P. De Filippi. 2018. *Blockchain and the Law: The Rule of Code*. Cambridge, MA: Harvard University Press.