Author accepted manuscript. Published under DOI 10.1108/JICES-12-2020-0128

Trust in Technology

Interlocking trust concepts for privacy respecting video surveillance

Sebastian Weydner-Volkmann (Ruhr-University Bochum)
Linus Feiten (University of Freiburg)

This is an author accepted manuscript (postprint) of the article published as:

Weydner-Volkmann, Sebastian; Linus Feiten (2021): "Trust in technology: interlocking trust concepts for privacy respecting video surveillance". In: Journal of Information, Communication and Ethics in Society 19 (4): 506-520. DOI: 10.1108/JICES-12-2020-0128.

Keywords: trust, philosophy of technology, IT-security, ethics, surveillance, privacy

Abstract

Purpose

We defend the notion of "trust in technology" against the philosophical view that this concept is misled and unsuitable for ethical evaluation. In contrast, we show that "trustworthy technology" addresses a critical societal need in the digital age as it is inclusive of IT-security risks not only from a technical, but also from a public layperson perspective.

Approach

From an interdisciplinary perspective between philosophy and IT-security, we discuss a potential instantiation of a "trustworthy information and communication technology (ICT)": a solution for privacy respecting video surveillance. Here, strong data protection measures address grave concerns such as the threat of bulk biometric tracking of citizens. In a logical argument, however, we show that this technical notion of "trust" needs to be complemented by interlocking trust relations in order to justify public trust.

Findings

Based on this argument, we demonstrate that the philosophical position considering "trust in technology" to denote either "reliability" or "interpersonal trust" is too limited as it fails to address critical aspects of IT-security. In a broader, socio-technical sense, however, we show that several distinct accounts of trust – technical, interpersonal and institutional – should meaningfully interlock, in order to address concerns with ICTs.

Value

This conceptual study demonstrates the potential of "trust in technology" for a more comprehensive evaluation of ICTs within the context of operation. Furthermore, it adds to the discussion of trust in IT-security by highlighting the layperson's challenge of judging a technology's trustworthiness. Vice versa, it contributes to Ethics of Technology by highlighting crucial IT-security needs.

1. Introduction

With its prominent introduction in the context of the EU's strategy on trustworthy AI applications (AI HLEG 2019), the concept of "trust" has moved to the center of attention for the ethical and political debates around information and communication technologies (ICTs). This comes after the last decades have already seen "trust" to strongly gain relevance in many debates inside and outside of Philosophy (Hartmann 2010; McLeod 2020; Searle, Nienaber, and Sitkin 2018). While this seems like a promising complementary development, the notion "trust in technology" was met with strong reservations by prominent philosophers of trust:

As Hartmann (2010, 15–16) states, many of their contributions start with a conceptualization of what is to be understood as "trust". This has resulted in a range of diverse definitions with partly conflicting premises, and hence, as Hartmann concludes, we cannot assume that it is possible to find one definition for trust that resolves these conflicts and integrates the different accounts.[i] Nevertheless, at least for normative conceptions of "trust", Harald Köhl (2001, 131) argues that the original domain of trust phenomena is that of *interpersonal relations*. Hartmann (2010, 20) reinforces this by claiming that no convincing conceptualization of "trust" has been made that is not characterized by interpersonal traits. For him, talking of "trust in technology", is therefore highly problematic and either turns out to imply *trusting engineers and designers*, or denoting a different phenomenon altogether: *the reliability of technical systems*. This view was concurred with by Thomas Metzinger (2019), one of the experts that drafted the guidelines for trustworthy AI mentioned above: in an interview, he described the idea of trustworthy technologies as "conceptual nonsense [orig. *begrifflicher Unsinn*]", which, he fears, is nothing more than a marketing term promoted by the industry used for some kind of ethics washing.

Despite the prominence of philosophical reservations to "trust in technology" (cf. also McLeod 2020), the last decades have seen a persistent interest in this topic. This is especially true for innovative medical technologies (Nickel and Frank 2020, 375; Nickel 2011), nano-technologies (Weckert and Sadjad 2020; Stebbing 2009) and digital technologies (Ess 2020; Coeckelbergh 2012;

Cap 2015). Furthermore, "trusted hardware", "roots of trust" or "chains of trust" have become established concepts in IT-security (Tehranipoor and Wang 2012; GlobalPlatform 2018) – concepts that have come to play a crucial role in digitalized societies.

In addition, recent trends in the philosophy of trust have, opened up new avenues to engage with "trust in technology", especially in interdisciplinary discussions: some philosophers suggest to move away from starting the analysis with one conceptual definition of trust. Instead, it is suggested to make parallel use of thin and thick conceptions of trust (McLeod 2020), to think equally about what would qualify as trustworthy in a certain applicatory context (Budnik 2016; Jones 2012) or even to *deliberately construct* a conception of trust depending on the purpose of the inquiry at hand: we should "ask not what our concept is, but what it ought to be, if we wish the notion to do useful conceptual work. Proposed accounts are to be evaluated against our legitimate purposes, which opens up the possibility that different accounts might suit different purposes" (Jones 2020, 7). This raises our main research question: how can "trust in technology" be conceptualized in order to play a meaningful role for the ethical evaluation of technology?

As an answer, we demonstrate that "trust in technology" allows us to address a critical societal need in the digital age that cannot be met by referring to interpersonal trust in the designers of an ICT or by assessing its technical reliability. ICTs also need to be trustworthy in the sense that they are robust against malicious actors that may try to use ICTs to deceive us for their own goals. While established conceptions of "trust" in IT-security may justify such a robustness for technical experts, however, "trustworthy technology" allows us to bring forth the perspective of those *who actually become vulnerable to attacks on or misuse of ICTs within a context of operation* and who are, more often than not, laypersons.

We will begin our logical argument in Section 2 by introducing a situated example for a class of ICTs that are particularly problematic with regard to trust relationships: digitized public video surveillance (VS). After a brief introduction of core reasons for public concern (Section 2.1), we will present a potential instantiation of a corresponding "trustworthy ICT" in Section 2.2, which is based

on previous work (Feiten et al. 2016). Ultimately, the aim of the underlying technical concept (called the "Digital Cloak of Invisibility", DCI) can be broadly understood as warranting trust in VS by addressing prominent ethical and societal concerns raised for conventional as well as for "smart" or "connected" VS. For this paper, we will limit our analysis to the role of trust with respect to the DCI's privacy and data protection mechanisms. This will provide a situated context that is necessary to continue in Section 3 with a discussion of the implied trust relations in terms of IT-security in contrast to the two alternatives proposed in the philosophical discussion for trust in technology: technical reliability (Section 3.1) and interpersonal trust in the engineers (Section 3.2). Based on the philosophical literature, we will then discuss further distinct trust relations in Section 4, mainly from an interpersonal and institutional perspective, that need to meaningfully interlock with these technical concepts in order to justify considering the DCI a "trustworthy technology" within its context of operation. Section 5 will then conclude the paper by discussing the broader implications for research on "trust in technology".

2. The Digital Cloak of Invisibility as a trustworthy ICT

2.1. Concerns with video surveillance

The use of camera systems for VS (also called "closed circuit television", CCTV) promises to address many public security issues such as identifying suspects as part of a criminal investigation, displacing criminal activities from places under surveillance, or enabling cost-effective real-time monitoring of places with few personnel. Apart from the legal admissibility of using VS systems for such purposes, however, ethical concerns have been raised especially regarding the proliferation of surveillance systems in public places. In particular, the continuous recording of multiple public places has raised concerns of also enabling the tracking of persons as they move through surveilled areas, which can then allow far reaching conclusions about the private lives and personal contacts of specifiable individuals. To raise public awareness, VS systems were awarded the negative German Big Brother Award (2005) "for the creeping degradation of citizens to objects of surveillance and the act of playing down dangerous tendencies towards ubiquitous observation".

Since ever more aspects in our lives involve automated, digital processing of personal and sensitive data in one form or another, however, *digital* forms of surveillance have immensely gained in relevance: Big data applications have made the bulk collection and processing of personal data a feasible and cost-effective undertaking (Snowden 2019). In this context of looming mass surveillance, the introduction of new systems can turn out to be just as problematic as the augmentation of more mundane surveillance technologies, like older CCTV systems, with ever more capable digitized functionalities: Today, we face the reality of AI driven technologies generating biometric templates for large parts of a population on behalf of state actors as in China (Mozur 2018) or on behalf of private companies in the US (Hill 2020) or in Europe (Schieb 2020). This enables the re-identification and tracking of persons on a very large scale, both in live video feeds (even from older, existing camera systems) and in previously recorded, stored data. Data protection and oversight measures that once seemed to suffice may, hence, cease to do so today.

Given this reality, there is ample reason to worry about misuse and disproportionate impact of VS in public places, a worry that might be rendered in terms of distrust in public VS. In the last decades, this distrust has caused considerable political frictions (e.g. Dachwitz 2018). Apart from the immediate impact of such technologies, researchers have also pointed towards so-called chilling effects: if citizens are concerned that public surveillance data might be misused, this might contribute to them refraining from perfectly legitimate or even socially desirable actions out of fear that the recordings could be used against them at a later point in time (Grimm, Keber, and Zöllner 2019, 39). What is particularly problematic from a trust perspective with regard to (digitized) VS is that the persons subjected to surveillance generally have little to no choice regarding the use of specific technology solutions, manufacturers or service partners. This makes (digitized) VS a particularly challenging context for grounding trust relations and a good situated example for the development of our logical argument.

2.2. *Privacy-respecting video surveillance*

In previous work (Feiten et al. 2016), we have described a socio-technical concept, called the "Digital Cloak of Invisibility" (DCI), which aims at minimizing undesired effects of widespread VS as well as at limiting the potential for misuse. This concept is based on a camera (as an instantiation) that offers privacy preserving functionalities, but that is part of a larger concept of operation meant to alleviate the grounds for public mistrust described above. Its foundational idea is a variant of the four-eyes principle, where both parties are institutionally divided and mutually independent – something not unlike a separation of powers, where the police require a warrant to search a citizen's home. With DCI-enabled surveillance, individual-related surveillance data can only be viewed if authorized by an independent third party. Figure 1 gives a visualization of this process.

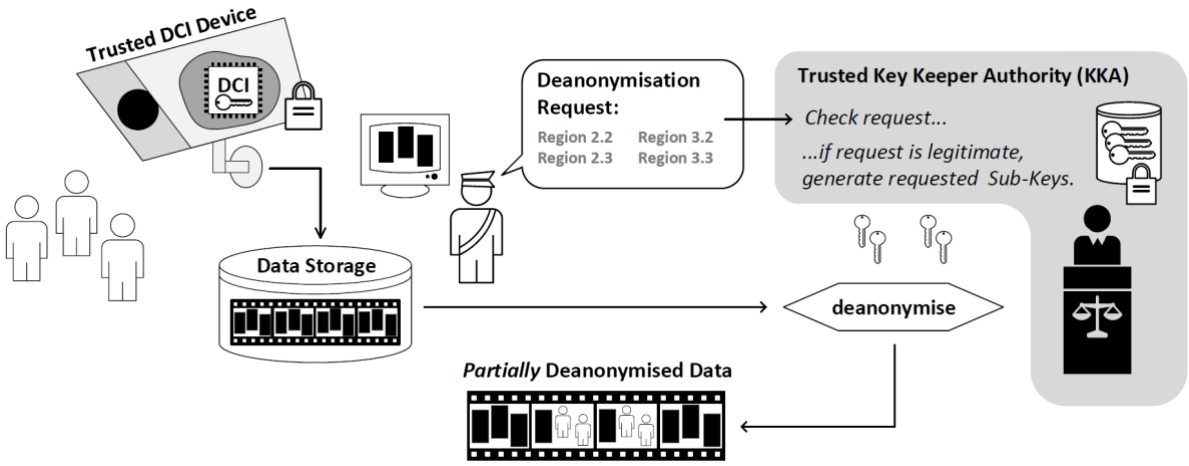

Figure 1: The concept of revocable anonymization through the Digital Cloak of Invisibility (DCI)

Within the camera runs an algorithm that identifies individual-related information (i.e. image regions showing people) and obfuscates it. Depending on the specific requirements of the context of operation, the technical capabilities of this algorithm may be very simple (e.g. obfuscating everything that moves in a mostly static environment) or more sophisticated (selectively detect and obfuscate people, number plates, etc.). These capabilities are to be chosen in such a way that the camera only outputs data in which all individual-related information has been *anonymized*. The non-individual-related data remains viewable (e.g. the image background and contours of anonymized people within it). Thus, many goals of real-time surveillance are still possible, i.e. a human observer can still

recognize dangerous situations (e.g. violent movements or mass panics). For stored data, the viewable non-individual-related data can also be used to identify which persons need to be *de-anonymized* in case their identity must be obtained, e.g. as part of a criminal investigation.

To perform such a de-anonymization, the approval of a third party (called the Key Keeper Authority, KKA) is required. Which individuals the KKA is comprised of (judges, civil ombudspersons, works council representatives, etc.) can, again, vary depending on the context of operation. The important factor is that the KKA itself has no direct access to the recorded (and anonymized) data and is not suspect of secretly colluding with those who have it.

On a technological level, the anonymization is made possible by the camera not actually deleting the individual-related data in the obfuscation process, but instead encrypting it with a key that is unique to every camera and securely stored within it. For every piece of individual-related information (i.e. every image region in every frame), the camera derives a sub-key from this master-key. The KKA knows the master-key of the camera and can thus also derive every sub-key given the parameters (frame number and region coordinates). This allows for a *selective* de-anonymization; personal data of people not relevant for the request (e.g. not part of an investigation) can remain anonymized.

In principle, the possibility to use surveillance footage *ex post* as evidence in a criminal investigation or even for digital biometric analysis remains unchanged by the DCI. All the KKA has to do is to issue all the relevant sub-keys required to de-anonymize the surveillance data in question. However, the KKA has the ability to granularly restrict the time span of this video data, i.e. the impact of a de-anonymization could be minimized depending on what is strictly necessary for a particular legitimate purpose. Furthermore, since this has to be done for each camera by one (centralized) KKA or potentially several independent (decentralized) KKAs, the DCI system would effectively protect the general public from misuse such as the mass biometric analysis of stored (or stolen) VS recordings from different cameras. Hardware “hacks” that would allow the extraction of the master-key can be hampered sufficiently with methods of “trusted hardware” in IT-security. While a successful attack can never be fully ruled out, established methods allow raising the costs of an attack to a level, where

it can be assumed to be highly unprofitable and very unlikely, especially if such attacks need to be repeated for each camera or for several (decentralized) KKAs.

Technologically a functional implementation of a DCI infrastructure is already feasible and a description thereof can be found in Feiten (forthcoming). We therefore posit that DCI-enabled VS alleviates the grounds for public mistrust in VS. As we will see, however, even if the instantiation is sound and well implemented as a socio-technical system within its context of operation, those technical aspects are not yet sufficient for the DCI to be a "trustworthy technology". In order to show this, we will first need to discuss the technical trust concepts implied by the DCI and show that this goes beyond questions of reliability and interpersonal trust.

3. IT-Security's concept of trust

3.1. "Trust in technology" beyond reliability

In the last section, we have referred to the concept of "trust" as it is understood in IT-security research. As mentioned in the introduction, some philosophers would criticize this use of the term "trust" and suggest using "reliability", instead (Hartmann 2010, 15–16; Metzinger 2019). Following the idea of a productive conceptual construction of "trust in technology", one may wonder if what makes the DCI a "trustworthy VS" solution could sufficiently be explained when referring to the concept of reliability. As will become clear, however, "trust" in the context of human-machine-interaction allows addressing additional issues, essential to the digitized world, that cannot be adequately reflected by "reliability". To demonstrate this, we first look at how the two terms are used in IT-security research.

In computer science, the term "reliability" has long been understood as the probability of a system to work correctly up to some expected time of failure. It is about the question of *when* – not *if* – a system will fail. Although there is no mechanical abrasion in microchips (as opposed to other machines that may fail due to this), effects like electromigration or dielectric breakdown can ultimately lead to a chip dying of old age. Besides such persistent defects of the hardware, there are also transient faults ("soft errors"), which only produce temporary malfunctions. Causes for these can

be for example "terrestrial cosmic rays" hitting a microchip (Ziegler 1996). Defects caused by aging are seen as system-inherent. Transient faults may have system-external causes, but since these causes can never be perfectly excluded, they are regarded as part of the system's normal operating conditions and hence as system-inherent as well. In computer science, the term "reliability" is used to describe robustness against such *system-inherent* failures in terms of probability. For a DCI-enabled camera, "reliability" would assess how likely it is that a camera breaks or malfunctions of its own accord during normal operating conditions.

In contrast to this, the term "trust" is used along with the concept of "security" (Tehranipoor and Wang 2012). While "reliability" is about robustness against system-inherent failures, "security" – and hence "trust" – are about robustness against deliberate attacks by a (system external) malicious actor. In some cases, the technical countermeasures against such deliberate attacks are similar to precautions against transient errors. There are, for example, *fault attacks* where the attacker aims to induce transient errors into the chip (Bar-El et al. 2006). Making a chip robust against fault attacks can therefore be similar to making it robust against system-inherent errors (Feiten et al. 2015). However, the crucial difference is that system-inherent errors *occur anywhere with a given probability*, whereas a deliberate fault-attack aims to induce errors at exactly the right time and location. Calling a chip "trusted" (e.g. secure against fault attacks) has therefore a different implication than calling it "reliable" (against randomly occurring errors): in the former case, the engineer makes considerations about possible weak points, where an attacker could strike. In the latter case, the engineer makes considerations about random error probabilities over time.

Furthermore, the kinds of damages resulting from reliability failures are generally different from those resulting from trust/security-infringing attacks. When reliability fails, the system usually malfunctions in an observable way or just stops working altogether. In terms of trust/security, however, denying a system's availability in a noticeable way (denial of service) is only one of many possible attack goals. There are just as many cases where the attacker aims at compromising the confidentiality or integrity of the data stored or processed by the attacked computer system – ideally

without being noticed. If a DCI-enabled camera is called "trusted" in this sense, it means that it has been designed to prevent any attack one would deem worth its goal *from an attacker's point of view*, e.g. extracting unencrypted plain data or the secret encryption keys. While absolute security is not possible, the estimated costs of a successful attack can be increased to a point where one would expect no attacker to invest the corresponding effort. This shows the relation between "trust" and "security": no device is 100% secure, but it is trusted if the effort needed for a successful attack has been increased enough to make it sufficiently unlikely.

Another difference between "reliability" and "trust/security" is that there are sometimes opposing requirements for both (Rahman, Forte, and Tehranipoor 2016). An example is the security application *Physical Unclonable Functions*, where a unique signature is generated from a chip's physical characteristics (Rührmair, Sölter, and Sehnke 2009). This allows to embed a secret in a chip which cannot be read out as easily as regular memory, and it allows to unambiguously identify a chip. For this to work, the physical characteristics of each chip must be slightly different, which can never be fully avoided in the production process. So this variation is a benefit for security, but from a reliability point of view, the uncertainty introduced by it is highly undesirable. Similarly, the ageing of chips is one of the main concerns of reliability, whereas the security method "hardware metering" takes advantage of aging to identify counterfeited chips (Dogan, Forte, and Tehranipoor 2014). These contradicting requirements for "reliability" and "trust" show that in computer science, the two concepts are clearly distinct and used for different, partly conflicting purposes.

3.2. "Trust in technology" and interpersonal concepts

Outside of this technical distinction between "reliability" and "security/trust", we argue that Hartmann's alternative suggestion – i.e. that "trust in technology" may denote trust in the engineers – does not adequately address the idea of security against attacks, either. This would imply the personal integrity of designers not to *deliberately implement backdoors or malicious functions* for their own gain. We will return to this problematic idea. Even then, however, this would not make a device "trusted", as it would not account for the modern fact that somewhere in our interconnected world, a

range of malicious actors – from professional competitors to state sponsored hackers – constantly attack computer systems.

This reality cannot be adequately captured in terms of interpersonal relations of trust or distrust, because many of these relationships are distinctly *impersonal*: I may never know the person who encrypted the research files on my computer and, in fact, that attacker may also never know me as a person, just as the owner of a computer system she is trying to extort bitcoins from (in exchange for giving me back access to my research files). Hence, when I wonder about security/trust with regard to my personal computer system, I can only refer to a general, undirected distrust founded on the knowledge that there are attackers that will spend *some*, but probably not vast amounts of effort on attacking my system.

In order to speak of a trustworthy system, this general, undirected distrust needs to be accounted for and one of the established ways to do this is following IT-security methods during design. They are intended to make computer systems *hardened against attacks* up to a level that justifies "putting trust" in their functioning *to my interests* rather that someone else's. One may trust any surveillance camera upon receiving it from a trusted designer (one may even design it oneself); but in order for it to be a "trustworthy camera" in daily operation, it needs to continuously warrant this trust because attacks have been made unreasonably "expensive".

One may argue, however, that technical "security/trust" (and technical "reliability") can be put in terms of "trust in the engineer" if this implies trust in the *professional competence and care of the engineers to have designed a secure (and reliable) system*. Trust in the engineer would then actually denote three *discreet* ideas: care and competences of (1) designing for technical reliability and (2) designing for technical security/trust as well as (3) the personal integrity of the designers. Regarding the latter, however, the actual designers are commonly unknown to those who do the trusting. Even individually authored smaller programs make use of APIs, open source libraries, compiler applications etc. that involve the work of many third parties. This makes it virtually impossible to determine who needs to be trusted. Furthermore, this trust becomes especially hard to justify in the surveillance

context where well-known companies and law-enforcement operators have abused trust and acted deceptively in the past. As an alternative to an overburdened concept of interpersonal trust, we suggest the idea of interlocking several distinct technical, interpersonal and institutional trust concepts. In the next section, we demonstrate this conceptual interlocking for the DCI system and we introduce further conceptions of trust required for a "trustworthy camera".

4. Interlocking trust concepts in human-machine interactions

In the previous section, we have focused on a technical discussion of *reliability* and *trust/security* for digital systems. For complex systems like a DCI-enabled surveillance camera, however, "trustworthy technology" should include further forms of trust relationships when seen in its context of operation. This is because even in cases where the technology functions reliably and securely on a technical level, some of the intended positive effects of increased privacy and data protection may still fail to materialize: many of those under surveillance lack an expert understanding of the concept and its implementation; if they refused to place trust in the concept and its implementation, the DCI system would fail to ease the political tensions around VS measures and it would fail to diminish corresponding individual concerns that may lead to chilling effects.

Hence, in addition to (1) forms of trust that relate to the technical soundness of the concept and its implementation, there are two further aspects to be analyzed: It needs to be shown (2) that there is an actual manifestation of trust amongst the surveilled persons in operational contexts; and (3) that placing trust in the DCI system is *epistemically warranted also for those non-experts*. Aspect (2) will not be the focus of this paper; it is planned to be answered as part of an empirical study in the future. Instead, the remainder of this paper will focus on aspect (3). Only when all three aspects meaningfully interlock can we speak of a "trust in technology" in a comprehensive sense as part of an ethical assessment: It then means (1) that, to an *expert assessment*, the technical concept is sound; (2) that we can show that the DCI system *empirically* supports the creation of trust in the socio-political context of non-experts; and (3) that it does so *for the right rather than the wrong reasons.*

This last aspect is important to highlight, because there are many wrong, i.e. imprudent reasons to place trust in technology, especially in surveillance technology. If we want to distinguish unwarranted from warranted trust in this sense ("trustworthiness"), we must therefore presuppose an *active rational component* in judging whether to place or refuse to place trust (cf. O'Neill 2010, 11–14). This means that, if the motivations to trust a technology are based more on *deception* than on *informed reasoning and autonomous consent*, then no matter how sound the technical concept behind the technology is, it still wouldn't be rationally warranted to *place* trust, but rather to *withdraw* it (O'Neill 2010, 85–87).

4.1. Trusting expert reviewers

How, then, can we describe such "prudent reasons" for non-experts to place or refuse to place trust in ICTs based the DCI system? As Donick (2019, 7–11) notes, one established way to warrant trust could be to make the technical concept, the involved code and the implementation procedure available for public review ("open hard- and software"). However, in most applicatory contexts, asking those under surveillance to competently judge the soundness of the technical concept and implementation for themselves is highly unrealistic. Furthermore, simply disclosing the source code of a DCI-enabled camera cannot prove that cameras in the field are, in fact, running the publicized code. Onora O'Neill (2010; 2017) addresses a similar point, when she argues that the mere fact of availability of information is not enough for warranting trust. It only implies *a theoretical possibility* to replace trust with an expert degree confidence for anyone determined enough to review the code and the hardware concept.

A system like the DCI therefore needs more than just an open-source implementation: There needs to be some kind of *mediated trust* that involves specifiable individuals whose duty and responsibility it is to conduct an expert assessment of the technical concept. In order to act as a kind of mediating link in the chain of trust between surveilled non-experts and the DCI system, the expert reviewers then need to provide a report that is commonly intelligible. Going a step further, some public applications may require for intermediary experts to witness and vouch for the process of a surveillance device being put into operation, as this is the only time these intermediary experts can verify what source

code the device is actually programmed with. Methods from hardware security and cryptography allow them to embed their signatures into the approved code and bind the compiled code to one device only. This device can henceforth cryptographically prove at any point that it is still running its original code and that it has not been replaced or manipulated (Feiten forthcoming).

The people under surveillance, in turn, then need to have good grounds for placing trust in these intermediary experts. Following O'Neill (2017), this requires the means to assess the competence, honesty and reliability of those intermediary experts. One important building block for this is the ability to identify the expert reviewers by name. This can help to assess if they have a good professional reputation among peer experts that speaks to their competence – a thought that O'Neill (2017; 2010, 58) models after academic practices. While this may suffice to judge competence, the notoriously hard issue to assess the honesty and reliability of actors remains – which brings us to the heart of the philosophical debate around interpersonal trust relationships. In her influential paper "Trust and Antitrust", Annette Baier (1986, 259) defines trust as the "reliance on others' competence and willingness to look after, rather than harm, the things one cares about and which are entrusted to their care." This includes a certain *discretionary power* by the trusted party regarding *how to care for what was entrusted* (Baier 1986, 240). What those under surveillance entrust to the intermediary expert is, according to this, primarily the technical assessment and implementation of the DCI system and the discretion on how to best do this.

In a mediated sense of interlocking trust concepts, however, what is entrusted by them is the protection of fundamental rights like privacy and data protection. This implies that in trusting the DCI concept due to an expert review, "one is necessarily vulnerable to the limits of that good will" of the expert reviewer and, therefore, "reasonable trust will require good grounds for such confidence in another's good will…" (Baier 1986, 234). One of the best grounds for trusting, Baier (1986, 243) argues, exists when the entrusted goods are *shared goods*, i.e. goods that both parties of the trust relationship individually care for (cf. also Koehn 2003, 5).

Following this argument, for a public application of the DCI, some governmental agencies typically tasked with IT-security responsibilities may not be a good choice for conducting such an expert review. The issue, here, is not that one may suspect the agencies to act against the interests of the public. Rather, in many instances, the agencies may serve *conflicting interests* and may, thus, not be tasked *unequivocally for the furthering of those specific goods that are to be entrusted in the review*. For the DCI, this makes it preferable, from a trust perspective, to task reviewers from a civil society background that have a longstanding and unequivocal reputation for the furthering of privacy and data protection – a reputation that would be further strengthened through conducting honest and reliable reviews).

An additional link to strengthen trust, complementing reputational aspects, could be established by making the act of entrusting the review public, in the sense of expressing confidence in the moral ability of the expert reviewer(s) to care for those goods (Rooney 2010, 347). The assumption for this dynamic is that placing trust in someone can, in itself, become a reinforcing factor for that person to behave worthy of that trust, i.e. that there is a certain "creativity of trust". As Philip Pettit (1995, 212f) argues, this dynamic is less dependent on the moral integrity of the trustee, but rather on their self-interest: being considered trustworthy generally implies to be seen in well regard – a good that most people strive for, especially when it is recognized by others: "… the existence of independent witnesses to the act of trust will provide further regard-centered motives for them to perform as expected. Let the trustor down and not only will they lose the good opinion that the trustor has displayed or promised; they will also lose the good opinion and the high status that the trustor may have won for them among third parties" (Pettit 1995, 215).

4.2. Trusting institutions: the Key-Keeper-Authority

This dynamic of "trust-responsiveness", in which a trustor has "reason to trust someone, even when he actually has no reason to believe in the other's pre-existing trustworthiness" (Pettit 1995, 216) can also be leveraged to provide good reasons on behalf of the public to place trust in arrangements like the KKA, which is a crucial part for the DCI's trust concept. For *public institutions*, however, this can

only work if certain conditions are met, Pettit (1995, 207, 202) warns; otherwise "we are in danger of designing institutions that will reduce trust or even drive it out."

Some of these conditions apply to the general character of society: it must be realistic for institutions to act independently (Pettit 1995, 223f) and there must also be a certain minimal level of social cohesion, as well as some examples for trustworthy civic engagement. After all, if every political act is seen along partisan lines or if cynicism regarding any form of political engagement abounds, the dynamic of trust responsiveness cannot find a suitable basis for an institution like the KKA (Pettit 1995, 221–23).

As long as the civic and institutional basis for democracy is functioning reasonably well, however, institutional arrangements can be created that allow trust-responsiveness to unfold and serve as a link to trusting a DCI-enabled surveillance camera. Pettit (1995, 224; cf. also Baier 1986, 240) argues that for this, trustees like the KKA must have meaningful discretion in deciding when to hand over the decryption keys and when to refuse. If the KKA's decisions were strictly controlled by sanctions, trust-responsiveness could not unfold even if the KKA, indeed, acted worthy of trust: "The more likely explanation of the manifestation … will always be that the trustor expects the trustee to be motivated by the sanctions" (Pettit 1995, 224).

O'Neill (2010, 57–59) argues in a similar way when she criticizes that we should not follow conceptions of trust that leave little room for such discretion: bureaucratic and legal means meant to introduce control in a trust relationship may effectively damage the latter. Instead, it must be possible for the trustor to judge in some way if the trustee acted in alignment of the entrusted good (Pettit 1995, 220f). It would be naïve for those under surveillance to place trust in the KKA if there was no discernable way to learn if they acted trustworthy or not; and trust-responsiveness couldn't unfold if the KKA would know that untrustworthy behavior would go unnoticed.

Hence, there have to be reports on the number and nature of instances of decryption to the public. For such contexts, O'Neill's (2010, 32) conception of "intelligent accountability" emphasizes the need to specify the duties and obligations – something she calls the business end of rights such as privacy

and data protection. With regard to the danger of misuse, ensuring that surveillance data is only shared in measured quantities and in connection to demonstrable cases of investigation would be the KKA's duty. In addition to reports on how the KKA fulfilled this duty, at least in some contexts, its handling of the decryption keys would also become publicly known as part of criminal trials and could be judged accordingly. Over time, some KKAs would gain a reputation as being trustworthy, others to the contrary (here, decentralized KKAs significantly limit the impact).

4.3. Trust, distrust and mediated trust in surveillance contexts

As discussed above, "institutional trust-responsiveness" requires some aspects of a well-functioning democracy as a basis. If this is so, however, it begs the question why there would be the need for such an elaborate socio-technical system as the DCI in the first place: if we are assuming a well-functioning democracy anyway, why not simply consider it prudent to place trust in the police and other law enforcement institutions, instead?

If surveys on trust in public institutions are any indication, trust in the police differs greatly across European societies.[ii] But even for countries like Finland or Germany, the fact that there is a high rate of trust in the police does not answer the question of whether we have *good reasons* to trust them with access to masses of public surveillance data: Just as argued above, even though the police may generally be considered to act in the best interest of the public, it may still be doubted whether there is an unequivocal alignment of the entrusted goods of privacy and data protection. From our trusting in many situations does not follow that we should entrust *everything in every context* (cf. O'Neill 2010, 9). This is especially true regarding the proliferation of surveillance ICTs.

One argument for distrusting law enforcement institutions to have relatively easy access to surveillance data concerns the protection of this data. In order to be trustworthy in the sense established above, these institutions would need to plausibly ensure that access to the data is limited and based on lawful, proportionate grounds. This competence would entail that we would need a good reason to assume that *all individuals with access or authority to request such data* will act competently, honestly and reliably – or at least that abuse of this access will be seldom enough to be

largely negligible. Given the recurrent scandals on data misuse by police officers, however, such a trust may be seen as naïve.[iii] It is safe to assume that there will always be *some* bad actors in law enforcement, in addition to criminals and organized crime that would go to great lengths to get access to such data through extortion or hacking as well as domestic and foreign secret services with a strong interest in such data. Thus, since we cannot assume blanket trustworthiness for an institution like the police (or for surveillance operators in general) and since we cannot assume an unequivocal alignment of the entrusted good of privacy and data protection, it would be imprudent to trust the police *as a monolithic institution* to have access to troves of VS data – unless there are strong technical and organizational protections in place. As we have seen in the previous sections, the DCI system can meet exactly this demand and form the basis to warrant entrusting access to VS data to the police.

Hence, instead of reading the deployment of a DCI system as an outward sign of *public distrust in the police*, one could, in fact, see the DCI system as providing a *reason to place trust not blindly, but with good judgement*: As part of the implemented DCI system, the KKA can also act as a mediating link for trust in police access to VS data by ensuring that legitimate, proportionate grounds exist. As a socio-technical system, the DCI's enforcement of limited, granularly adjustable access to the ever growing masses of surveillance data with an equally growing potential for misuse, hence, effectively prevents scenarios of mass surveillance and blanket biometric tracking of citizens in live or recorded video feeds that cause mistrust in VS.

This is only true, however, insofar as we can point out good, context-specific grounds for the public to place trust in the DCI. More generally, the notion of "trustworthy technology" can be used to address the perspective of laypersons (e.g. the public) by situationally interlocking technical with institutional and interpersonal trust concepts. Such interlocking trust concepts can also be inclusive of attacks on and misuse of ICTs – crucial issues, especially for surveillance contexts, that may be overlooked if "trust in technology" is purely put into terms of reliability or trust in engineers.

5. Conclusions

In this paper, we have started from the observation of a growing relevance of the idea of "trust in technology" in the technical, political but also philosophical debates. However, prominent philosophers have voiced reservations towards this idea and suggested either using the term "reliability" or rendering trustworthy technology in interpersonal terms. In order to defend the notion of "trust in technology", we built on newer trends in the philosophy of trust that involve the productive conceptualization of "trust". From an interdisciplinary perspective between philosophy and IT-security, we looked at a specific candidate that could count as a "trustworthy ICT": the Digital Cloak of Invisibility (DCI). The DCI is a concept for privacy respecting video surveillance (VS) with strong data protection measures that enforce limited, granularly adjustable access to surveillance data. Thereby, the DCI system addresses the growing potential for misuse, like bulk biometric tracking of citizens in live or recorded video feeds, which gives reasonable grounds for mistrust in VS and may cause chilling effects within the public.

Instead of developing one single account of trust for our analysis, we demonstrated that several, interlocking accounts, technical, institutional and interpersonal, can be put to productive use to describe different crucial aspects of public VS. After a brief introduction of the DCI and the concerns it is meant to address, we showed that neither interpersonal accounts, nor the term "reliability" suffice to include all relevant technical aspects: both fail to address potentials for misuse and attack of IT-systems by malicious actors – issues that have gained crucial social and political relevance and that are of particular concern for "smart", interconnected surveillance technologies. For this, IT-security research has developed technical trust concepts that need to be included in a more comprehensive ethical analysis.

However, as we have subsequently shown based on the DCI, such methods of trusted hard- and software are not enough to justify the idea of "trustworthy technology" in a broader, socio-political sense. The DCI, but also ICTs in general, need to be carefully arranged within their social context of operation: several distinct forms of trust relationships need to meaningfully interlock, in order to comprehensively address "trust in technology". Only then can ethics (and an ethics-sensitive

implementation strategy) be inclusive of the situational problem of assessing a technology's trustworthiness from a layperson perspective.

6. Literature

AI HLEG. 2019. "Ethics Guidelines for Trustworthy AI." Bruxelles: High Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Baier, Annette. 1986. "Trust and Antitrust." *Ethics* 96 (2): 231–60. https://doi.org/10.1086/292745.

Bar-El, Hagai, Hamid Choukri, David Naccache, Michael Tunstall, and Claire Whelan. 2006. "The Sorcerer's Apprentice Guide to Fault Attacks." Edited by Fawwaz T. Ulaby and James E. Brittain. *Proceedings of the IEEE* 94 (2): 370–82. https://doi.org/10.1109/JPROC.2005.862424.

BigBrotherAwards.de. 2005. "Technology: Video Surveillance." Big Brother Awards. 2005. https://bigbrotherawards.de/en/2005/technology-video-surveillance.

Budnik, Christian. 2016. "Gründe für Vertrauen, Vertrauenswürdigkeit und Kompetenz." *Deutsche Zeitschrift für Philosophie* 64 (1): 103–18. https://doi.org/10.1515/dzph-2016-0007.

Cap, Clemens. 2015. "Kann Man Einem Computer Vertrauen?" In *Vertrauen*, edited by Josette Baer and Wolfgang Rother, 109–26. Basel: Schwabe.

Coeckelbergh, Mark. 2012. "Can We Trust Robots?" *Ethics and Information Technology* 14 (1): 53–60. https://doi.org/10.1007/s10676-011-9279-1.

Dachwitz, Ingo. 2018. "Überwachungstest am Südkreuz: Geschönte Ergebnisse und vage Zukunftspläne." *netzpolitik.org*. October 16, 2018. https://netzpolitik.org/2018/ueberwachungstest-am-suedkreuz-geschoente-ergebnisse-und-vage-zukunftsplaene/.

Dogan, Halit, Domenic Forte, and Mark Mohammad Tehranipoor. 2014. "Aging Analysis for Recycled FPGA Detection." *2014 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, October, 171–76. https://doi.org/10.1109/DFT.2014.6962099.

Donick, Mario. 2019. *Die Unschuld der Maschinen: Technikvertrauen in einer smarten Welt*. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-24471-2.

Ess, Charles M. 2020. "Trust in Information and Communication Technologies." In *The Routledge Handbook of Trust and Philosophy*, edited by Judith Simon, 405–20. New York: Routledge.

Eurobarometer. 2019. "PublicOpinion - European Commission." https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Chart/getChart/themeKy/18/groupKy/88.

Feiten, Linus. Forthcoming. *Take the Power Back! Secrecy, Accountability and Trust in the Digital Age.* Dissertation zur Erlangung des Doktorgrades der Technischen Fakultät der Albert-Ludwigs-Universität Freiburg.

Feiten, Linus, Matthias Sauer, Tobias Schubert, Victor Tomashevich, Ilia Polian, and Bernd Becker. 2015. "Formal Vulnerability Analysis of Security Components." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34 (8): 1358–69. https://doi.org/10.1109/TCAD.2015.2448687.

Feiten, Linus, Sebastian Sester, Christian Zimmermann, Sebastian Volkmann, Laura Wehle, and Bernd Becker. 2016. "Revocable Anonymisation in Video Surveillance: A 'Digital Cloak of Invisibility.'" In *Technology and Intimacy: Choice or Coercion*, edited by David Kreps, Gordon Fletcher, and Marie Griffiths, 474:314–27. Cham: Springer. https://doi.org/10.1007/978-3-319-44805-3_25.

Flade, Florian. 2020. "'NSU 2.0'-Drohbriefe: Polizisten in Hamburg und Berlin befragt." tagesschau.de. September 6, 2020. https://www.tagesschau.de/investigativ/wdr/nsu-zwei-punkt-null-101.html.

GlobalPlatform. 2018. "Root of Trust Definitions and Requirements v1.1." GP_REQ_025. https://globalplatform.org/wp-

content/uploads/2018/07/GP_RoT_Definitions_and_Requirements_v1.1_PublicRelease-2018-06-28.pdf.

Grimm, Petra, Tobias Keber, and Oliver Zöllner, eds. 2019. *Digitale Ethik Leben in vernetzten Welten.* Kompaktwissen XL 15240. Stuttgart: Reclam.

Hartmann, Martin. 2010. "Die Komplexität des Vertrauens." In *Vertrauen - zwischen sozialem Kitt und der Senkung von Transaktionskosten*, edited by Matthias Maring, 15–25. Karlsruhe: KIT Scientific Publishing. https://doi.org/10.26530/OAPEN_422381.

Hill, Kashmir. 2020. "The Secretive Company That Might End Privacy as We Know It." *The New York Times*, February 10, 2020. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

Jones, Karen. 2012. "Trustworthiness." *Ethics* 123 (1): 61–85. https://doi.org/10.1086/667838.

———. 2020. "Trust." In *International Encyclopedia of Ethics*, 1–9. Wiley. https://doi.org/10.1002/9781444367072.wbiee665.pub2.

Koehn, Daryl. 2003. "The Nature of and Conditions for Online Trust." *Journal of Business Ethics* 43: 3–19.

Köhl, Harald. 2001. "Vertrauen als zentraler Moralbegriff?" In *Vertrauen: die Grundlage des sozialen Zusammenhalts*, edited by Martin Hartmann and Claus Offe, 114–40. Theorie und Gesellschaft 50. Frankfurt/Main: Campus.

Maass, Dave. 2016. "Police Around the Country Regularly Abuse Law Enforcement Databases." Electronic Frontier Foundation. September 28, 2016. https://www.eff.org/de/deeplinks/2016/09/police-around-country-regularly-abuse-law-enforcement-databases.

McLeod, Carolyn. 2020. "Trust." In *The Stanford Encyclopedia of Philosophy*, edited by Edward

N. Zalta, Fall 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entriesrust/.

Metzinger, Thomas. 2019. "Nehmt der Industrie die Ethik weg! EU-Ethikrichtlinien für Künstliche Intelligenz." Tagesspiegel.de. April 8, 2019. https://www.tagesspiegel.de/politik/eu-ethikrichtlinien-fuer-kuenstliche-intelligenz-nehmt-der-industrie-die-ethik-weg/24195388.html.

Mozur, Paul. 2018. "Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras." *The New York Times*, July 8, 2018. https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html.

Nickel, Philip J. 2011. "Ethics in E-Trust and e-Trustworthiness: The Case of Direct Computer-Patient Interfaces." *Ethics and Information Technology* 13 (4): 355–63. https://doi.org/10.1007/s10676-011-9271-9.

Nickel, Philip J., and Lily Frank. 2020. "Trust in Medicine." In *The Routledge Handbook of Trust and Philosophy*, edited by Judith Simon, 367–77. New York: Routledge. http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781134881604.

O'Neill, Onara. 2010. *A Question of Trust*. 5. printing. The BBC Reith Lectures 2002. Cambridge: Cambridge Univ. Press.

———. 2017. "Linking Trust to Trustworthiness." Presented at the UCD Ulysses Medal award, University College Dublin. https://www.youtube.com/watch?v=A0u76tA1OyA.

Pettit, Philip. 1995. "The Cunning of Trust." *Philosophy and Public Affairs* 24 (3): 202–25.

Rahman, Fahim, Domenic Forte, and Mark Mohammad Tehranipoor. 2016. "Reliability vs. Security: Challenges and Opportunities for Developing Reliable and Secure Integrated Circuits." *2016 IEEE International Reliability Physics Symposium (IRPS)*, April, 4C-6-1-4C-6–10. https://doi.org/10.1109/IRPS.2016.7574542.

Rooney, Tonya. 2010. "Trusting Children: How Do Surveillance Technologies Alter a Child's Experience of Trust, Risk and Responsibility?" *Surveillance & Society* 7 (3/4): 344–55. https://doi.org/10.24908/ss.v7i3/4.4160.

Rührmair, Ulrich, Jan Sölter, and Frank Sehnke. 2009. "On the Foundations of Physical Unclonable Functions." *Cryptology EPrint Archive* Report 2009/277: 20.

Schieb, Jörg. 2020. "Siebter Himmel für Stalker: PimEyes ist ein Staubsauger für Fotos." *Digitalistan* (blog). July 14, 2020. https://blog.wdr.de/digitalistan/wie-pimeyes-unser-gesicht-ueberall-im-netz-entdeckt/.

Searle, Rosalind, Ann-Marie Ingrid Nienaber, and Sim Sitkin, eds. 2018. *The Routledge Companion to Trust*. Routledge Companions in Business, Management and Accounting. New York: Routledge.

Simon, Judith, ed. 2020. *The Routledge Handbook of Trust and Philosophy*. New York, NY: Routledge. https://www.taylorfrancis.com/books/9781315542294.

Snowden, Edward. 2019. *Permanent Record.* New York, NY.

Stebbing, Margaret. 2009. "Avoiding the Trust Deficit: Public Engagement, Values, the Precautionary Principle and the Future of Nanotechnology." *Journal of Bioethical Inquiry* 6 (1): 37–48. https://doi.org/10.1007/s11673-009-9142-9.

Süddeutsche Zeitung. 2020. "Mehr als 400 Verfahren wegen Abfragen an Polizei-PCs." Süddeutsche.de. June 26, 2020. https://www.sueddeutsche.de/politik/abfragen-polizeicomputer-missbrauch-1.4979314.

Tehranipoor, Mohammad, and Cliff Wang, eds. 2012. *Introduction to Hardware Security and Trust*. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4419-8080-9.

Weckert, John, and Soltanzadeh Sadjad. 2020. "Trust in Nanotechnology." In *The Routledge*

Handbook of Trust and Philosophy, edited by Judith Simon, 391–404. New York: Routledge.

Ziegler, J. F. 1996. "Terrestrial Cosmic Rays." *IBM Journal of Research and Development* 40 (1): 19–39.

---

[i] The multidisciplinary literature on trust has grown far too complex to summarize in a paper introduction. For philosophy, recent revisions of encyclopedia entries (McLeod 2020; Jones 2020) and a handbook publication (Simon 2020) offer a very good overview.

[ii] In a late 2019 study, agreement with the statement "tend to trust the police" ranges from 37% (Albania) to 94% (Finland). In Germany, 85% state that they tend to trust the police (Eurobarometer 2019).

[iii] For example, in the recent "NSU 2.0" incidents, members of the German police have unlawfully accessed information on predominantly female public persons, who have subsequently received extreme right wing threats (Flade 2020). The extent of the involvement of police officers is under investigation in this case. It is clear, however, that the misuse of access to police databases is a more general problem not only in Germany, but across many democracies (Süddeutsche Zeitung 2020; Maass 2016).