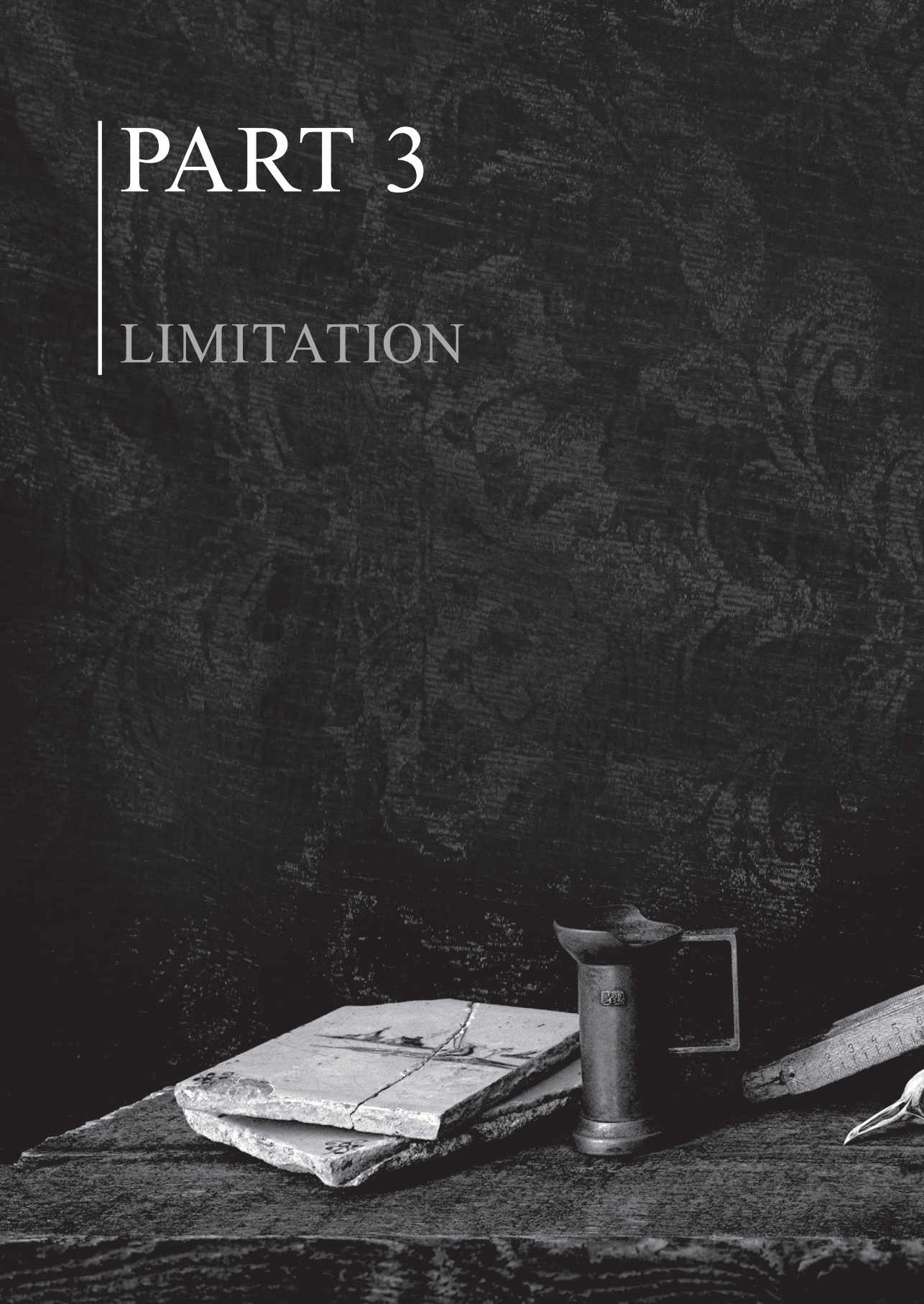
Computer says no: understanding digital authority

Weggemans, D.J.

PART 3

LIMITATION



7 HUMAN SAYS NO

Stanislav Petrov worked in *Serphukov-15* – a secret bunker southwest of Moscow and the nerve center of the Soviet Union's early-warning system for incoming ballistic missiles. His job was to monitor the system - code named Oko ("Eye") - and to directly report any alerts to his superiors. On 26 September 1983, just after midnight, alarms sounded. A red dot appeared on the electronic map covering the entire wall of the control room. It indicated that a missile was launched from an American base and was coming their way. "The siren howled very loudly, but I just sat there for a few seconds, staring at the screen with the word "LAUNCH" displayed in bold red letters," he told the BBC in 2013.[108] Not much later, the system picked up a second, a third, a fourth and a fifth nuclear missile. The red letters changed to "MISSILE STRIKE". The computers double-checked the data and displayed the reinforcing additive: "HIGH RELIABILITY". If this was true, the missiles would detonate on Soviet territory within 20 minutes.

Petrov had a choice. He could do what the Soviet military protocol dictated and report the incoming missiles to his superior. It was most likely a retaliatory strike would then follow, causing millions of deaths.[109] He could also declare the incoming missiles a false alarm and prevent a large-scale nuclear war, but with the risk that the Soviet Union would be critically hit if the missiles were real. Years later, Petrov described the choice he was facing:

> "My cozy armchair felt like a red-hot frying pan and my legs went limp. I felt like I couldn't even stand up. That's how nervous I was when I was taking this decision" (Barton, 2011).

Far from being certain, he decided to refute the computer's output and told his superior that there had been a system malfunction. A 50-50 guess he admitted (BBC, 2013b). "Got it", the officer - who, according to some accounts, was drunk at the time[110] – replied, and the message went up the chain (Hoffman, 2009: 11).

We know now that Petrov made the right call. After twenty-three minutes he realised that nothing had happened: "if there had been a real strike, then I would already have known, it was such a relief" (BBC, 2013b). It was found that the satellite had mistakenly

taken sunlight's reflection on a rare high-altitude cloud formation above North Dakota for a missile launch. "The sky always holds surprises for us," Petrov concluded (Spiegel, 2010). The algorithm for filtering out such data clearly needed revision and Petrov became known as the "man who saved the world".[111]

Doing what a digital technology tells us is important for the successful functioning of many things in our world – including safety and security. But it comes with serious problems when taken too far. As we have seen throughout this book, under the sway of digital authority people can be drawn well beyond the lines of reason and persistently follow technology into disaster. Fortunately, we do not always listen to these technologies. In fact, telling a computer "no" does not have to be difficult and it seems that many of us are all too familiar with situations in which we refuse to do or believe what the thing says. Still, while doing otherwise can be simple in theory, it might not be easy in practice. Especially when the stakes get higher and information scarcer, it seems an incredibly hard thing to do for everyone – as the story of Stanislav Petrov also illustrates.

The moment people do not accept the output of a computer signals the absence of authority and is illustrative of the idea that like human beings, digital technologies are not authorities by default, but only are when recognised as such by others. This chapter, the third part of the book, is about the limits of digital authority. In paragraph 7.1, I describe what it means when people say no to a digital technology. The following paragraphs, in turn, focus again on the relational, situational, and cultural layers of authority. The premise here is that the absence of authority can be traced back to elements in relations, situations, or cultures that obscure the construction of digital authority. Paragraph 7.2 argues, for example, that it is the absence of asymmetry in the relation between a technology and a human being that forms an explanation for people's dismissive attitudes and responses towards digital authority claims. In paragraph 7.3, I consider the use of different situational definitions and scripts that lead people to not recognise digital authorities. The cultural grounds for refuting digital authority are discussed, in turn, in paragraph 7.4. In the following section, paragraph 7.5, I focus on the moments in which digital authorities are challenged. What can digital technologies do to secure compliance among their human subjects in times of resistance? Paragraph 7.6 concludes this chapter with a discussion on how human beings and societies can be equipped to deal with the risk of erroneous or false digital authorities and how to mitigate their potential impact. Ultimately, dealing with digital authorities remains mainly a question of when one should rely on their communications and when to say no. The final paragraph presents a summary of part 3.

7.1 "I WILL NOT DO WHAT YOU TELL ME!"

The counter-intuitive and sometimes spooky things we can do under the sway of authority are clearly fascinating, and to date still much research is focused on authorities' potentially destructive speech. Among other things, the paradigm of human obedience to authority was transferred to the domain of human- computer interaction to test how far a person would go in torturing a robot (Rosalia et al, 2005; Bartneck & Hu, 2008). Still, people often do not listen to those claiming authority and may, for whatever reason, choose to think or act otherwise. And this holds true for many situations, including Milgram's experiments - in which a substantial part of the participants, at some point, refused to do as they were told (Miller, 1990; Packer, 2008) – as well as for our interactions with digital technologies.

The moment a message is questioned and fails to gain compliance constitutes the absence of authority. A quick observation of many of our daily interactions reveals that examples hereof are plenty. Not every teacher effortlessly silences a class, not every police officer can make a suspect cooperate without force, just like not every digital technology will be simply listened to by its users. Of course, drivers sometimes do end up in ditches when blindly relying on their GPS, but people also frequently ignore the digital navigational voices – with mixed success. The same will go for those working in a security setting; not every security official will automatically follow the recommendation of just any digital technology all the time.[112] Apparently users are not left to the whims of technology.

When someone rejects digital authority by saying "I do not believe it" or "I will not do it" it can reflect two things. First, it can indicate that authority was never there in the first place. For example, people might structurally ignore the output of a technology from the moment it gets implemented. Second, an audiences' refusal can also signal the negation of digital authority at a later stage. Here it is perhaps better to speak of the *deconstruction* of authority. People can, for example, at some point in time, come to question or dispute a digital technology's claim. Authority comes with great expectations, and when a technology fails to live up to these it will quickly disappear. While it is conceptually interesting to further concentrate on the failed construction and deconstruction of authority, for the purpose of this book I prefer a more general focus on the thing both have in common: the struggle of a digital technology to be consequential.

This rejection of digital authority, in turn, can manifest itself in various ways. It is most visible when people stop using a technology altogether. They will not even turn it on. Or they can simply end up ignoring it – the output of the digital system is not considered in human decision-making processes at all. This can be done *actively* (a technology's output is deliberately bypassed or openly refuted) or *passively* (the technology's output is not registered by the decision-maker in the first place). Digital communications can also be used in a more selective way. That is, individuals rely on their own observations and reasoning and only use the technology as an instrument for justification towards others. All these responses reflect a general rejection of digital authority.

Furthermore, rejecting, or challenging digital authority can happen at different levels too. It can be an individual response in which users become irresponsive to what a technology tells them. But it can also happen more collectively when peers start to openly call the technology into doubt. Such responses can be signaled through visible behavior revealing human defiance or in open verbal justifications for why the technology's communications must not be followed. In these situations, others rejecting digital authority can inspire individuals to do the same.

Last, as argued in chapters 2 and 4, sometimes it is (almost) impossible to work around or challenge digital systems effectively. Someone might try to say no or to reject their influence, but eventually some technologies will work their "will" regardless of what human subjects want. The code does not allow for it, the individual is not granted access to change the outcome. Still, when people do try to resist a digital technology, it signals the absence of authority. As a reminder, when people are forced to do as a technology tells them it is not authority but coercion. The rejection of authority it is not so much about whether people can act differently than a technology tells them, but more about their willingness to unquestionably rely on what they are digitally told.

Having established what it means to reject digital authority claims, let's now turn to *what* makes people not recognise digital authority. What makes us say or think: "no" when we interact with a digital technology? To answer this question, I draw focus again on the three layers constitutive of digital authority: the relation, the situation, and the culture in which needs and expectations of authority are shaped. When these layers do not coincide, digital authority claims will not resonate among its subjects. Below I investigate each layer to further understand such deviations.

7.2 RELATION: INDIVIDUAL AND TECHNOLOGICAL CHARACTERISTICS

Digital authority implies an asymmetrical relation between human beings and a digital technology in which the person feels the technology has a right to be listened to. This asymmetry is fundamental for all authority relations as it gives rise to trust, respect, and the willingness to accept whatever is being communicated. One way to understand the rejection of authority claims, in turn, is to consider it a consequence of the absence of such a relational asymmetry. Let's therefore take a closer look at the actors involved. Are there certain characteristics in people and technologies that frustrate the construction of such a hierarchical social structure between them? The following sections focus on the things that secure experiences of equality or even human superiority, in the relationships between people and technologies.

7.2.1 A matter of personality?

Psychological research on human authority has often focused on the personality traits of its subjects. And, indeed, it seems to make sense that people who easily accept and follow the words of proclaimed authorities have different personalities than those who do not. However, things are not so straightforward. One of the biggest insights from the obedience experiments was that the personalities of the obedient and disobedient participants were shockingly similar (pun intended). Elms and Milgram (1966), for instance, studied a variety of personality characteristics associated with obedience and defiance towards human authorities. They found very few differences between both groups. An exception was that disobedient participants scored significantly higher on "social responsibility".[113] It is plausible to think that those who have a low sense of social responsibility are more prone to obey an authority figure and shock an emotive victim. Conversely, someone who does feel responsible for what happens in his presence can reasonably be expected, at some point, to rebel against the one making certain claims or demands. Elms and Milgram (1966: 284-6) also found that authoritarian attitudes could predict obedience and disobedience. The participants that were fully obedient in the experiments consistently appeared to have more prior respect for authoritarian values than their disobedient peers. In other words, the research found a link found between peoples' authoritarian tendencies or personalities – as Theodor Adorno and colleague's (1950) called it - and their willingness to obey. This too is not much of a surprise, as also Bocchiaro and Zimbardo (2008: 30) observed: in a social situation with a clear hierarchical structure in which a lab-coated authority figure commands a research subject, you can expect that those with more respect for authoritarian values are most likely to do as they are told.

Elsewhere, the influence of moral reasoning on obedience in the Milgram experiments was discussed (Blass, 1991: 403). It was argued that with the moral development of individuals the likelihood of disobedience increased. And others also investigated dispositional factors such as social intelligence (Burley & McGuiness, 1977) and beliefs about the relationship between actions and outcomes – the extent to which one thinks one has personal control over one's life or that this is controlled by external forces such as fate or divine influence (Blass, 1991: 405). However, like with most other variables that have been tested, the outcomes were contested and deemed unreliable (Bocchiaro & Zimbardo, 2008: 31).[114]

What these psychological analyses suggest is that personalities are generally not the best predictors for individuals' decisions to challenge or refute authority. While it is intuitively attractive to assume that disobedience or non-compliance stems from a very particular set of personality factors there is not a lot of objective evidence for most such reasoning.

But what does this mean for our discussion here on the denial of digital authority and the relations between human beings and technologies? After all, the scope of our investigation does not just concern actors in formal positions of authority that explicitly command or order within an experimental setting but also include those that advice or suggest and operate in many everyday situations that could lead people to respond in

accepting or rejecting ways. Yet, studies into the role of personalities and the willingness to challenge technologies offer similar general insights. For example, Biros, Fields and Grunch (2003) found that people who are not optimistic about digital technologies in general are also less willing to rely on the output of a specific digital technology – a combat drone in their study. Szalma and Taylor (2011) found a negative correlation between neuroticism and agreement with a digital system (Szalma and Taylor, 2011: 82). Others found that introverts are less likely to listen to a computer than extraverts (Merritt & Ilgen, 2008). And Mcbride and colleagues (2012) showed that more intuitive nurses were more likely to accept the diagnoses from a digital expert system than nurses with a sensing personality.

But what this strand of literature also emphasises, above all, is that *personalities* are not a good predictor for human reliance on digital technology's communications. Other individual characteristics seem more relevant here. As discussed in Hoff and Bashir's (2015) systematic literature review on the issue of human-automation trust, *age* is often seen as a significant factor. But it differs in what way. Some studies have shown that younger adults, by default, trust digital decision aids less, but others, however, found that the more critical attitudes actually come with age (Hoff & Bashir, 2015: 414). Interesting in the context of this book is the study of Pak and colleagues (2012) who revealed that a picture of a physician in the interface of a digital diabetes decision support aid significantly enhanced younger participants' trust in the system's advice, but not that of older participants. Last, Hoff and Bashir (ibid) mention the role of *gender* in trust and reliance on digital technologies. But here too, researchers found no gender-based structural differences.[115]

These studies provide some interesting suggestions on the individual characteristics that could explain the differences in people's tendency to accept or reject digital authority. At the same time, this is exactly what they are: suggestions. Lack of research, scope and significant findings do not lead us to conclude that there is a typical user that tends to refute or ignore digital output and claims to authority.

7.2.2 Ascribed authority: Judging the system

There are all kinds of grounds on which people can feel equal to or better than a technology, and thus most likely to reject claims of digital authority. One is the institutionally ascribed authority of those interacting with the technology. Let me illustrate this with an example from the security domain of a digital technology that failed to acquire authority.

The use of digital risk assessment technologies in criminal justice is rapidly expanding across the world. According to Sonja Starr (2015: 205) we have entered the risk assessment era. Digital risk assessment here is about the use of algorithms that calculate "risk scores" for criminal defendants that estimate the chance that a person will commit crime in the future based on statistical inputs such as age, gender, criminal record, employment status and education level (Stevenson, 2018: 304). There is substantial support for the idea that these technologies are better at predicting future arrest than people and

are therefore increasingly used for determining restrictions on liberty – for example in decisions about length of a sentence, probation supervision levels and parole conditions.

In 2011 a law was passed in Kentucky making it mandatory for judges facing pretrial custody decisions to consult an algorithmic risk assessment tool that estimates the chance that a criminal defendant would flee or re-offend before his trial. Deciding about whether to detain or release someone while awaiting trial is a critical step in the justice process and getting it right is paramount for both the defendant and the community-at-large. The use of the risk assessment system was expected to make judges better at assessing who is safe to release (Mamallan, 2011:3). The effective release of low- and moderate-risk defendants would dramatically reduce incarceration rates without affecting public safety.

While the system was supposed to make Kentucky's criminal justice system fairer and cheaper this is not what happened. Judges refused to listen to the technology. After the law came into force, there was only a small increase in the number of suspects sent home. Most of the time judges simply decided to overrule the system's recommendations. If they would have deferred to the digital system, then non-financial release would have been granted to 90% of the defendants. But, in practice, as Megan Stevenson (2018: 373) found, this only happened in less than a third of the cases. In fact, it did not take long for judges to return to their previous ways and after a couple of years the release rate was even lower than before the law took effect.[116] And the story does not end here. Stevenson also discovered that after judges were required to consider the digital risk assessment system's output, white defendants were allowed to await trial at home without case bail much more often than black defendants.[117] In a different study using Kentucky data, Alex Albright (2019) showed that this was not simply a consequence of different risk scores by race. Rather, judges were more likely to override the system's default recommendation and to impose harsher bond conditions for black defendants than for similar white defendants.

The case of Kentucky confirms the idea that there is no such thing as authority by default. For Kentucky, digital technology was clearly not the magic bullet to improve the criminal justice system as it lacked the ability to produce consequential communications. Lawmakers seemed to know this already for some time. Before the 2011 law, the use of the digital risk assessment system was optional and most judges dismissed it (Stevenson, 2018: 343). But maybe, they reckoned, the system just needed a bit of help in establishing its authority? By making consideration of the technology's output mandatory, the new law was supposed to support the digital claim to authority by guiding the judges' attention to whatever the technology had to say. The support turned out to be insufficient though. Not digital authority but human discretion and bias kept guiding human decision-making.

But why is this so? In chapter 4 we have seen that digital systems could easily be recognised as authorities in contexts in which alternative sources of information are lacking (epistemic niches) – something that also seems to apply to the decision-making of Kentucky judges. There is only so much they can know about the risk of flight or future crime and

thus consistently doing what the computer says would have made perfect sense. Moreover, reliance on the technology, would have allowed them to work efficiently and to avoid the discomfort of (felt) responsibility (see also paragraph 8.4). When an inmate would reoffend, they could always point at the wisdom of the technology. Hence, why did the Kentucky judges did recognise the digital risk assessment tool's authority?

Stevenson (2018: 369-70) suggests that one of the reasons why judges only selectively incorporated digital risk assessments into their bail practices is that some might not have believed that these systems provided new and useful information: they did not recognise the system's expertise and thus did not experience the fundamental inequality of their relationship vis à vis the technology. This idea is also supported by a survey in which judicial attitudes to the provision of risk assessment technologies were investigated (Chanenson & Hyatt, 2016: 10-11). The survey indicated a general belief among judges that their judgement was more accurate than that of a risk assessment tool. In fact, less than 10% of the responding judges believed that the digital systems were indeed better at predicting future crime than themselves (ibid). The following section discusses the role of one's own perceived expertise in more detail.

Another ground for why these judges did not recognise the technology's authority has to do with their own position and the nature of the judicial office in general. As put by one author: "no other official is held to the standards of impartiality required of a judge" (Bedner, 2002: 14). Impartiality is fundamental to the judicial office and if judges fail to maintain it, their legitimacy will be damaged. Therefore, they are typically trained and frequently reminded to think for themselves, to make autonomous decisions. It is also anchored in their authoritative positions and the discretionary room granted to them. This authority in the executive sense (paragraph 3.2), translates into the relative autonomy to make decisions, using individual judgement or assessment when determinate rules are absent. As a Supreme Court order explicitly stated: "nothing in these guidelines shall be construed to limit the court's discretion as to whether or not to grant pretrial release to a defendant" (Supreme Court of Kentucky, 2011: §11). The result was that judges, indeed, frequently relied on their own authority to ignore the system's advice. The focus on their own authoritative role thus can also be a factor that prevents individuals' willingness to recognise digital authority.

The Kentucky case illustrates that real-world implementation of digital technologies can be different from what one might expect. Even when used in a context of limited information and time, and when its use is made mandatory by law, information systems can thus fail to gain compliance. When unrecognised by human authorities, digital systems remain technologies. Not authorities.

7.2.3 Expertise and system knowledge: wiser than computers

Authority relations often hinge on a sense of differentiated access: the idea that someone has relevant knowledge or experience from which oneself is barred. So, what if people come to hold such knowledge or insights themselves? Contrary to the static or enduring personality traits I discussed above (e.g., age, personality, gender), *domain-* and *system knowledge* are dynamic characteristics that people develop over time. Both are reasons for rejecting digital authority claims as they challenge the potential hierarchical structure of the human-technology relationship.

Domain knowledge or expertise is one of the most important reasons to say no to a digital technology. Expertise here refers to knowledge a person acquires over time. It is the result of extended experience in one specific domain. With greater expertise also often comes greater self-esteem or confidence, which is typically associated with a stronger tendency to rely on one's own judgement and resistance to influence attempts of others (Sagarin & Wood, 2007: 329). Research has shown that people with greater domain expertise are less likely to depend on digital technologies than novice users do (Sanchez, et al, 2014).[118] And it is also not difficult to come up with daily life situations in which our expertise leads us to refute the output of a technology. Think for example of drivers ignoring their GPS navigation in the area they live in because they know a shortcut or a medical specialist not needing an expert system to diagnose a patient because she has "seen this disease already a thousand times". Ergo, when people (think they) know things themselves they need the technology less and expertise provides a strong ground for such feelings.

In addition to expertise, we must also consider individuals' knowledge about a particular digital system. Having *system knowledge* is relevant for determining the limits and weaknesses of a digital system. It makes users aware of what a system can and cannot do and of its potential weak spots. Such knowledge can stimulate people to remain wary of unthinking compliance. As an illustration we can use Stanislav Petrov's life-or-death choice. Petrov "had a funny feeling in [his] gut" he told the Washington Post in 1999. The Soviet defenses were extensively briefed about a possible U.S. all-out nuclear attack – a single blow that would leave them completely paralysed. So, when five red dots appeared on the electronic map, doubts arose. "When people start a war, they do not start it with only five missiles (..), you can do little damage with just five missiles," Petrov reasoned based on his military expertise he had built during his twenty-six years of military service (Washington Post, 1999). But just as important seemed his technical background and knowledge of the digital system. As described elsewhere, Petrov was more of an engineer than a soldier:

> "He liked the logic of writing formulas, often used English-based computer languages. On most days, he was not in the commander's chair but at a desk in a nearby building, working as an analyst, responding to glitches, fine-tuning the software. Only twice a month he took an operations shift in order to keep

> on top of the system. When Petrov arrived eleven years earlier, the station was new, with equipment still in crates and the room empty. Now it had grown into a bristling electronic nerve center" (Hoffman, 2009: 7).

As the deputy head of the combat algorithms department at Serpukhov-15 Petrov knew the system's algorithms like no other; he designed the software and operating manual himself. He also experienced how the system was rushed into service as a response to the United States' introduction of a similar system but that the Soviet system was still troubled: shortcomings will be fixed later they were told.[119] Still, the electronic boards showing a missile launch never happened before. One of the things that made him suspicious was the speed in which the data was processed through the 30 levels of verification after the target was identified; Petrov doubted if the system could do that. He also considered various potential system malfunctions that could explain the alert. As he later said in interview with the German magazine Der Spiegel (2010): "We are wiser than the computers. We have created them". His system knowledge surely contributed to him not following the computer's output in an automatic and unthinking fashion. Petrov has always remained humble every time he was called "the man who saved the world". "I just did my job", he told a journalist (BBC, 2013b), "but they were lucky it was me on shift that night and not a dulled soldier". The latter, according to Petrov, would have probably followed the protocol and the digital technology rather than reasoned for himself.

Knowing a technology or the domain in which it operates helps to transcend the mysterious air that is constructive of the asymmetry that is endemic to authority relations. It is important to maintain a critical attitude that allows users to question digital authority and say no. This also goes for past negative experiences one can have with a (similar) digital technology (Manzey et al., 2012). Past failures, errors, and other malfunctions can also be grouped under the system knowledge header; when we have experienced a technology failing, it provides users with relevant lessons about its weaknesses that might prevent blind or automatic deference. This brings us to the next section about the performance of a digital technology itself.

7.2.4 Technological failures and inconveniences

When a digital technology fails to live up to our expectations, we are likely to reject its output. This often relates to the objective performance of the digital technology. When a technology does not provide consistent output – the technology is unreliable – or does not do what it is claimed to do – the technology is invalid – it impacts how we relate to it. This also goes for frequent error messages and breakdowns (Merritt & Ilgen, 2008). When digital technologies fail to perform authoritatively in the eyes of their human audience – e.g., they are caught being wrong or unreliable - compliance becomes highly uncertain. It seems that for digital authorities the adage also holds true that first impressions matter. Failures that

are experienced in the early stages of interaction tend to have a greater negative impact on individuals' willingness to rely on a technology than errors that occur later (Manzey et al., 2012). And especially "easy failures" seem detrimental to authority. After all, why should someone blindly rely on a technology for a complex and important task when it cannot perform its basic functions well?[120] The argument goes that when digital technologies fail or break down, the structure of the relationship changes. People are confronted with the technology's fallibility and warned that automatic and unthinking reliance is unwarranted.

A digital system that offers no clear advantages but only seems to make a task more arduous will also not convince an audience to do as it is told (Parkes, 2012). Indeed, this also ties in with the previous paragraph on expertise and knowledge. When people realise that a technology "is nothing special" and following it is in no way beneficial to them or others but will only generate extra work, digital authority claims surely become problematic.

7.3 SITUATION: SIGNIFICANT OTHERS AND WRONG CUES

Digital authority lies vested in the character of the situation. When people enter a new situation, they face the question: what is going on and what is appropriate here? In chapter 5, I argued that the right situational cues can easily trigger people to define situations with digital technologies as authority situations in which accepting or submissive responses are warranted. When digital authority is rejected, in turn, it can be considered the opposite outcome of this situational analysis – the setting a person entered did not lead him or her to experience it as if he or she was dealing with a digital authority. By contrast, the individual determined something is not an authority situation in the first place or rather recognised the authority of another actor present. Below I discuss some of the potential mechanisms that link to such responses.

7.3.1 The wrong impression

One way to understand an actor's failure to be consequential is as a sign of the absence of relevant situational cues that suggest authority. When thinking of technologies, this could bring us to the issue of *design*. For example, it is suggested that when system interfaces look less attractive, people are less willing to rely on them (Hoff & Bashir, 2015: 422). Or think of situations where the wrong symbols are used. When the portrayed insignia, logos or other things intended to mark the authority of a technology are not recognised as such - for example because they have become controversial or do not match the expectations of an audience – they will not be able to yield authority effectively.

In chapter 5 the idea of anthropomorphism was also discussed. Humanness in digital technologies can provoke feelings such as trust and respect that are important for reliance on digital technologies (Pak et al., 2012; De Visser et al., 2012). It can even stimulate trust resilience in people – their trust does not decline as much when system

errors occur. However, when designers do not consider individuals' expectations of a technology's interface, the result can be quite different. As mentioned earlier, when done wrong, designing humanlike technologies produces uncanny feelings and will frustrate digital authority claims instead of supporting them (Mori, 1970; Nass & Gong, 1999). In fact, it can be argued that it is impossible to perfectly mimic human authorities, and that every attempt to do so is doomed to fail as it will essentially reveal the technologies "non-humanness" (Nass & Moon, 2000). This could remind people that they are interacting with an object, and that the unquestioning acceptance of technological output is silly. We realise that an authority script is not appropriate here and get snapped back to real-life as it were.

This also goes for the *experienced* performance of a technology. There are many studies that highlight the positive relationship between the usability of a technology and human compliance (Hoff & Bashir, 2015: 422-3). It was found that when humans know how to read or interact with a technology, they become more likely to rely on it. Wang, Jamieson and Hollands (2011), for instance, describe that users were more willing to follow a combat identification system when it presented its output more explicitly.[121] One way to interpret this is that such explicit communication forms an authority cue that incites automatic acceptance among human subjects. But when communications are vague or unclear, it might not predispose an audience to respond with such accepting or obedient attitudes. The situational cues emitted may not to be associated with authority and might therefore not trigger the relevant automatic scripts and situational definitions. When a technology gets to complex or its output to incomprehensible, it might only lead people to think for themselves.

More generally, research on human-computer interaction suggests that communication style is an important instrument for digital authorities. Digital technologies can be endowed with a wide variety of "personalities" and sometimes there is a mismatch between an artificial personality and its human user that will work against acquiring authority. Whereas "polite" (Nass, 2004; Spain & Madhavan, 2009), "patient" and "non-interruptive" (Parasuram & Miller, 2004) computers can be good in motivating users to accept their output (for example in medical diagnostics), other personality types can have the opposite effect (Nass & Lee, 2001). Therefore, with digital technologies, too, it is not just about *when* they perform (results), but also *how* they perform that makes whether they can shape an authority situation for us.

7.3.2 Significant others

The behavior of others present in the situation can strongly influence how we end up reading a situation and thus our potential recognition or rejection of digital authority. In his studies, Milgram also demonstrated the importance of the social environment for disobedient responses to authorities. For one of his experiments, he used two rebel peers (1974:116-121). Both were actors, instructed to disobey at some point. The moment the victim "receiving" the shocks began to protest, one of these actors announced that he no

longer wanted to participate. Some trials later, the second actor withdrew as well. All tasks were now assigned to the naïve participant, as the experimenter had repeatedly insisted that the experiment had to continue. It was found that only ten percent of the participants in this variation of the experiment continued to obey to the end while in the initial condition this was more than 60 percent. While more research is needed here, a similar pattern seems to exist when thinking about digital authority; when people observe or hear strong negative sentiments among their peers against a digital technology, they themselves are also likely experiencing them as well. This was for example found by Michael Workmann who focused on subjective norms that "reflect an individual's assessment of social influences from important others such as managers and coworkers, about performing a behavior" – e.g., "my colleagues think I should not do X" (Workmann, 2005:213). Workman found that someone's perception of significant others' tacit and explicit views about a digital system significantly shaped their willingness to listen to a digital technology themselves. That is, the words and actions of others shape how individuals define and respond to situations with digital technologies. The dismissive acts or thoughts of peers are cues that affect individuals' ideas of what is going on and what actions are appropriate. It could lead them to conclude that rather than unquestioning acceptance, it is deviance that is warranted here. When thinking about any form of authority, audiences always matter.

Moreover, the *salience of a subject* - the person or thing about whom a decision is taken - can also influence people's willingness to follow a digital technology. Those who are affected by output of a computer are often distant or out of sight completely. They remain names or numbers on a screen such as is the case with many databases or mere images with a warning sign. Digital technologies have the tendency to *neutralise* their human subjects. But those who are affected by a computer's output can also be right in front of us, or we can easily identify ourselves with them. Under these circumstances doing or thinking what the technology tells us can become quite difficult. This can also be derived from the Milgram experiments where it was found that individuals' willingness to follow an authority figure depends on how clearly the negative consequences are presented (1974: 34-55). In the experiment's baseline setting, the participants were sitting in an adjacent room. While the complaints of the victim could be heard through the wall, they could not see him. In one variation of the experiment, Milgram placed the victim only a few feet away, in the same room as his subjects of study. With victims that could now be heard and seen, the level of obedience dropped from 62,5 percent in the initial condition to 40 percent (ibid).[122] These findings suggest that it is easier to follow an authority when the victims are out of sight. It is easier to push a button to kill a distant target than to kill someone face to face. Distance allows people to ignore, or even forget, that those who might be negatively impacted by their actions are real human beings. Not only are individuals more likely to respond empathically when they see a fellow human being in distress, moving a subject closer also confronts them with the causality between their actions and the negative impact on another. People are more likely

to reject authority when they see themselves as personally causing the harm that follows from doing as they were told.[123] And while it is true that with digital technologies those who are affected often remain distant subjects, making the impact of a decision largely invisible, even in these situations victims can gain salience for example by overtly complaining or by making themselves visible to the human decision-maker. Digital neutralisation can be countered and subjects can incite human decision-makers to follow their own judgement instead that of a technology. It confronts the idea that one is in an authority situation and draws attention to the idea that it is in fact appropriate within that setting to rely on one's own judgement.

7.2.3 Whose authority?

Sometimes we enter a situation with multiple authority claimants. The question then becomes not so much whether we define something as an authority situation, but rather whose authority are we talking about? Especially when those involved end up making opposite claims this poses an important issue.

There are many situations in which human authorities disagree, such as when parents are making opposite demands to their children or experts disagree over some issue. People then face the choice who to listen to. Similarly, there can also be situations in which a human authority figure diametrically opposes a digital technology. Depending on the setting, people may very well decide or be actively motivated to listen to a human authority instead of a technological one.[124] And sometimes it is the presence of other digital technologies that forces people to choose between two potential digital authorities. Consider the following case from the field of security.

In 1988, with the end of the Iran-Iraq war nearing, the American Missile cruiser USS Vincennes shot down an Iranian passenger jet that passed the strait of Hormuz, between Iran and the United Arab Emirates. All 290 people on board of the Airbus A300 were killed, including 66 children. The cruiser was equipped with an *Aegis system* - one of the most sophisticated computerised defensive weapons systems used in combat at that time – which had wrongly identified the aircraft as an attack plane of the Iranian Air Force. While the radar systems of nearby warships had correctly identified the flight as civilian, they all decided to rely on Aegis' judgement which they recognised as superior. As one of the crewmembers of a nearby ship later explained: "We called her *Robocruiser*, she always seemed to have a picture" (c.f. Noorman 2023).

What happened with the USS Vincennes is that multiple systems provided diverging communications about what was the case. This output could by no means be checked by the military officials: human beings lack the senses to gather relevant information as well as the brain capacity to comprehend all the input they received under conditions of time pressure. They listened to the Aegis system as it was generally considered to be the most sophisticated and effective system to track and target incoming missiles and airplanes. Yet, there were some grounds for doubt. The Aegis system was designed to detect targets at

great range, but consequently was unable to distinguish the size of a target. This entailed that the USS Vincennes was best suited for patrolling an open sea environment. But while the cruiser was initially deployed to the Western Pacific and the Indian Ocean, the USS Vincennes was later dispatched to the Persian Gulf to patrol its confined waters during the Iran-Iraq war in the 1980s. In this environment, however, the lacking ability of the Aegis system to determine size became problematic and contributed to the disastrous outcome. As it was put in a *Newsweek* magazine article:

> "In combat on the open sea, size doesn't matter much; a small plane or missile can sink a ship. But in the Gulf, where big civilian planes share airspace with small warplanes, size can be a crucial clue." (c.f. Dotterway, 1992: 140).

With the ongoing digitisation of their world individuals will be increasingly confronted with competing authority claims made by different machines. Situations in which different technologies present competing authority claims raise the question what makes people recognise the authority of one technology over another. For example, is new always better? New technologies can also come with problems or negative associations – for example when an earlier technology is deemed unreliable it could also negatively impact expectations of other technologies that could perform the same task (Ross, 2008). There are still many interesting avenues for further research here. For now, let us conclude that within a particular setting, technologies sometimes end up competing for digital authority and that it is up to human beings to define the situation and decide which gets recognised as such.

7.4 CULTURE: PROGRAMMED RELUCTANCE AND UNCERTAINTY

Cultures shape people's ideas and expectations of authority. They condition expectations of authority and technologies and influence which digital authority claims resonate within a given audience or, by contrast, will be refuted. Several cultural grounds exist that could lead individuals to challenge a digital authority.

7.4.1 Hospitable cultures

In chapter 6 it was already discussed that while authority has a place in every culture, there are considerable differences in expectations towards authority and ideas about the effective exercise of authority. It was suggested that across cultures, those in authority are often performing in very different ways. In some cultures, where individuals are socialised into following hierarchies and recognising inequalities in social relations, a strong emphasis on official symbols and formal statuses might help to effectively yield authority. However, in other environments such strategies can be counterproductive, and a more egalitarian approach towards an audience could be more successful. Different cultures may come with

different ideas about what is authoritative and what is not. One way to consider the absence of digital authority, then, is as the consequence of a disjuncture between a technology's expressions and individuals' culturally conditioned expectations.

There is also research that specifically focuses on the influence of culture on human-computer interactions in less abstract terms, highlighting that in some places people are, in general, less inclined to rely on digital technologies in the first place. One study, for example, found that Americans were less willing to listen to a decision aid compared to Mexicans (Huerta, Glandan & Petrides, 2012, c.f. Hoff & Bashir, 2015: 413). Elsewhere, significant cultural differences were found in how people respond to social robots (Hoff & Bashir, 2015: 413). The idea that some cultures are more open to technological innovation and display a greater general willingness to listen to digital technologies was already discussed in the previous chapter. While it was claimed that digital authorities can be found everywhere and that most cultures are characterised by a certain ambiguity towards technological development, it may very well be true that some cultural contexts are more hospitable to the emergence of digital authority relationships than others. It would be interesting for future studies to further investigate this issue and learn more about cultural differences in saying no to digital technologies.

7.4.2 Cultural unease and scandals

People's hesitance to recognise digital authority can also be rooted in societal experiences with the downsides and risks of science and technology. Common examples are disasters such as in Chernobyl (1986) and Fukushima (2011) that have strongly shaped contemporary perceptions of nuclear technology. Similarly, environmental degradation and climate change have become strongly associated with modern polluting technologies. Technological innovation has given rise to security risks that can be both global in reach and global in scale (Western, 2016: 100). It contributes to a general climate of technological unease, making people wary of technological innovation and complicating the construction of digital authority across cultures.

Critical incidents and negative experiences can also have a more specific cultural impact. This was also suggested in an article on the risk of algorithmic bias (Alon-Barkat & Busuioc, 2022). Alon-Barkat and Busuioc investigated reliance on algorithmic advice. Their study was conducted in the direct aftermath of a major scandal in the Netherlands, involving public authorities' overreliance on an algorithm with discriminatory effects – the "childcare benefits scandal" (*toeslagenaffaire* in Dutch).[125] They argued that the scandal resulted in an enhanced awareness of discrimination and algorithmic biases and a growing reluctance among people in the Netherlands to trust algorithms in its aftermath. It suggests that scandals can have a learning effect in societies – longer or shorter term – and, as such, influence the potential of digital authority.

7.4.3 Reflexivity

In the previous chapter it was argued that we now live in a reflexive reality where people are constantly monitoring and revising their world. In this world where tradition is challenged and under the growing influence of science and technology, everything has become more open and complex. People will constantly seek reliable information of others. This transformation has generated new authorities. To navigate their way through daily life, individuals rely on all sorts of authorities, including digital technologies that make the world a bit less uncertain and a bit more predictable. But these new authorities, too, are susceptible to potential revision and change as individuals are constantly integrating new knowledge to decide who to trust or listen to. While guidance by tradition has largely been replaced by guidance by expertise, in modernity there is no such thing as unquestionable knowledge – as discussed in paragraph 6.5. It is the thing that is fundamental to scientific and technological progress and when people learn more about science, they learn more about the idea that claims to knowledge can always be challenged and changed. There will always be competing knowledge claims or imprecise predictions, stimulating further the continual reflexivity of individuals and institutions. Hence, technological authorities are also fragile, in the sense that they too can always be challenged and replaced.

7.4.4 Organisational norms

Organisations often come with a distinct set of ideas about what is important, how things work and norms that regulate practices and direct organisational members towards certain goals (Reason, 1997: 192; Cooper, 2000; 112). The culture of an organisation influences how people experience their work and interact with other people. This can also include conditioning people's ideas and expectations of (digital) authority. As argued in the previous chapter, the contingent and uncertain nature of modernity can lead organisations to obligate or socialise people in working with and listening to digital technologies. But the opposite can also be the case. Organisations can design all sorts of mechanisms to ensure responsible use of digital technologies and develop shared perceptions on organisational security. This for example includes implementing structures that facilitate independent thinking and reflection by professionals or by placing emphasis on human responsibility. As such, organisational cultures can both offer a ground for the construction of digital authority and for its deconstruction.

One recent study of Meijer, Lorenz and Wessels (2021) explored the use of predictive policing technologies in Berlin and Amsterdam. They found that in Berlin, the use of a digital system for the identification of crime hotspots was facilitated by "linking the acceptance of the predictive policing system as a valid source of information, to the promotion of the norm that usage of the information system is good police practice" (ibid: 841). Here, the higher organisational levels emphasised that the system needs to be used and portrayed it as a form of rationalising the organisation. Police officers stated that they

felt that resistance was futile, and that they had little alternative than to use the system (ibid). In contrast, in Amsterdam the technology was explicitly portrayed as a system that can certainly help police officers in their practices, but that these professionals must "seek interaction between their own insights and knowledge, and the information from [the technology]" (ibid: 843). In short, organisational norms for using a system can be quite different and can influence the extent to which people willingly rely on it.

7.5 ARTIFICIAL UNINTELLIGENCE

What can be done when digital authority is rejected? What happens when subjects reveal they do not want to listen and do not recognise a technologies' authority? There are several ways in which compliance can be ensured. A common initial response of human authorities is to resolutely repeat and reiterate what was said – potentially supplemented with a raised voice or a changed posture. This is what parents do ("I'm telling you to go to bed!") and what Stanley Milgram did when confronted with hesitant participants ("It is absolutely essential that we continue" and "You have no other choice" (Milgram, 1973: 63). And it is also the default response of most digital technologies when people do not comply. GPS-systems repeatedly tell drivers to turn around, just like a digital system might tell its users that someone is a risk every time a particular name is reentered. Reemphasising one's words or message can be a convincing strategy for reminding an audience of the unequal relationship between speaker and listener and to ensure compliance.

But sometimes more is needed (and sometimes it is really good that people do not listen). Some respond to resistance with statements that make them sound interesting or stimulating to listen to such as managers or politicians reminding an audience of their great achievements or future ideals. Others might start making compliments or aim to lead by example. When challenged authorities can also begin to explain themselves to convince an audience or remind people of the great benefits of compliance. Similarly, for people in positions of authority there is also often the option of coercing a person into doing something – a manager can fire, a police officer can arrest, and a parent can use physical force. Their powers can contribute to the reconstruction of authority – but it might just as well lead to the very opposite: a loss of trust and respect and authorities' definitive decay.

In other words, authority is dynamic and there are different ways to (re-)convince a hesitant audience. Recognising (the potential for) resistance then becomes an important factor for its maintenance. For authority to be sustainable an actor must be able to sense and respond to another's sentiments. Such social and intellectual skills to act wisely in social relations can grouped under the heading of *social intelligence* (Salovey & Mayer, 1990: 187).

A general definition of social intelligence is the "the ability to understand and manage people" (Thorndike and Stein, 1937: 275). Others have highlighted that such skills can also be directed towards oneself. Social intelligence, by extension, then includes both

the ability to perceive one's own as well as others' "internal states, motives and behaviors, and to act toward them on the basis of that information" Salovey and Mayer (1990: 187) write. They also observe that social intelligence is often described in a more manipulative manner - as the ability to steer or manipulate the responses of others or to get others consistently and voluntarily do certain things (ibid).

Dealing with feelings and emotions, in turn, forms an important aspect of social intelligence. In the interactions with our environment, we encounter loads of affective information that must be processed. Individuals differ in the extent to which they can do so. Salovoy and Mayer (1990:189) use the concept of *emotional intelligence* to describe the "ability to monitor one's own and others' feelings and emotions, to discriminate among them and to use this information to guide one's thinking and actions" (1990: 189). Like social intelligence, emotional intelligence can thus include both knowledge about the self and about others. In essence, emotional intelligence focuses on "the recognition and use of one's own and others' emotional states to solve problems and regulate behavior" (ibid).

Earlier, I discussed how authorities use a range of "theatrical instruments" to impress their audiences (paragraph 3.4). Such expressions help them to define the situation for others – they highlight that an actor is to be trusted and respected and should be listened to. Now, it could be argued that for effectively conveying and maintaining authoritative impressions social and emotional intelligence are relevant. Especially in times of hesitance or resistance, an authority must be able to sense an audience's dismissive attitudes and convincingly respond to them accordingly. That is, one must recognise emotions in others, regulate affect and use it to motivate subjects to behave in an accepting way. But how about digital authorities? Is this not a bit too much to ask from *things*? Can digital technologies recognise, express, and utilise emotions? If we accept feelings and emotions as relevant factors in maintaining authority relationships, then investigating the relationship between digital technologies and emotions deserves a bit more attention here.

In her famous book *Affective Computing*, Rosalind Picard focuses on "computing that relates to, arises from, or deliberately influences emotions" (1997: 3). She thinks that, for computers to be truly effective, they "will have to have emotions or emotion-like mechanisms working in concert with their rule-based systems" (ibid: 12). Picard distinguishes three general emotional abilities. First, there is the ability to *recognise* emotions – "to infer an emotional state from observations of emotional expressions and through reasoning about an emotion-gaining situation" (ibid: 50). This ability can require senses such as sight, hearing for gathering facial expressions, gestures, and vocal intonation (ibid). When speaking about computing technologies there are also other relevant inputs for recognising emotions that have no equivalent in human senses such as reading infrared temperature or measuring electrodermal responses (ibid). Second, there is the ability to *express* emotions. Emotions can be relatively easily perceived (such as via facial expression, voice intonation, gestures, movement, pupillary dilation, posture), or they can

be less apparent to others and require physical contact (such respiration, heart rate, pulse, temperature, blood pressure and muscle action potential) (Picard, 1997: 27). Picard argues that computers can express emotions without having them: "the basic requirement for a computer to have the ability to express emotions is that the machine have channels of communication such as voice or image, and an ability to communicate affective information over those channels" (1997:56).[126] Last, there is the ability to *have* emotions. Picard argues that feelings are often considered to be the factor that separates humans from machines (1997: 60). Surely, the ability to have emotions is the most complex and controversial when thinking about digital technologies and it remains to be seen if it will be ever possible - and desirable - to make machines feel.[127]

Picard then extends the concept of emotional intelligence to digital technologies. Building on the three abilities of affective computing that were mentioned above, she describes a technology with emotional intelligence as one that "is skilled at understanding and expressing its own emotions, recognising emotion in others, regulating affect, and using moods and emotions to motivate adaptive behaviors" (1997:76). For the argument here, it is the recognition, regulation and use of emotions in *others* that I am most interested in. That is, when thinking about authority it does not seem so relevant if, and how, digital technologies can recognise, regulate and utilise their own emotions for themselves, but rather the extent to which they could recognise, regulate and utilise emotions of others to get things done.

The work on affective computing is interesting because it explores the potential of digital technologies to manage and utilise emotional mechanisms (Picard, 1997:84). Fact is, however, that many digital technologies (still) lack such effective emotional abilities. Not only do they not have emotions, but most computers also lack the ability to perceive and express emotions that could help them lay claim (again) to the trust and respect of a hesitant audience. For digital technologies often goes that they only have limited access to the world of social emotions: they lack emotional intelligence. Sensing the right strategy to claim authority or responding to dismissive audiences, then, is inherently problematic. Computers are designed to respond in a certain way, still generally unable to improvise and manipulate the emotions of an audience to gain compliance. When people are in doubt or unwilling to recognise digital authority, there is only so much the technology can do. It can reiterate itself or warn a user. But it generally does not have many alternatives to tailor its expressions towards them for its words to remain authoritative. When digital authority goes unrecognised, the technology needs others (designers, managers, or lawmakers as was described above) to help them out.[128]

7.6 MAKING PEOPLE THINK

In an overly complex world, people need digital authorities – for security and many other areas. Oftentimes, digital technologies "know" what they are talking about which makes their directives of great value. But as we have seen throughout this book, human beings should always remain mindful as there is no guarantee that the technology is always right. Digital technologies, for whatever reason, can be wrong and automatic compliance may have severe consequences. Moreover, authority symbols can easily be faked. Digital technologies may look good and be advertised in slick commercials but sometimes turn out to be very different in practice. In these and other ways, blind deference could hinder the very thing on which digital authority's moral justification tends to rest: improved decision-making.

The central question, then, becomes how can people know when to listen to a digital authority and when to rely on their own judgement? According to Robert Cialdini (2007: 230) saying no to a human authority begins with posing the simple question "is this authority truly an expert?" With this question Cialdini shifts our attention to the authorities' credentials and the relevance of those credentials to the topic at hand. By focusing on the *evidence* for an authority status, Cialdini concludes, major pitfalls of automatic deference can be avoided (ibid). While for digital authority this question can also be a good general starting point, determining expertise of digital technologies is often quite problematic as most people lack the general skills and experience to do so. Especially since many digital technologies are implemented with the argument of improving human decisions or are introduced as "knowledge" or "expert" systems, there is often only limited initial evidence for people to focus on.

In interactions with digital technologies, the aim perhaps should not just be to gather evidence for an initial authority status, but to maintain a focus on the authoritative communications themselves and to stimulate active assessment of what we are being told. After all, it is an old philosophical mistake – going back at least to Plato – to suppose that even with the identification of the greatest expertise, unconditional or automatic compliance is always warranted. Paraphrasing Kant (1784), the question to ask here is: how can people be urged to also think for themselves under the influence of digital authority?

7.6.1 Enhancing human efficacy

It is easy to be amazed by all incredible digital qualities that are praised as superior. But focusing on all the things a technology can do can also make people wonder: who am I to disagree? What do I know? People could feel they do not have the *capacity* to critically assess or challenge a digital technology. This brings us to the notion of *efficacy*.

Efficacy broadly refers to having the ability to achieve a desired outcome. In the context of this book this entails having the competency, skills, knowledge, and other resources to make one's own decisions. Enhancing individual efficacy would not only increase the possibility that someone recognises erroneous digital output before it is

put to use, but it can also counter the sense of powerlessness that was described above. An enhanced sense of personal or *self-efficacy* – the belief in our own abilities to reach specific goals (Bandura, 1978) - could make individuals more confident to challenge an authority as they come to view themselves as having the capacities to do so. This was already argued by Kelman and Hamilton (1989), stating that enhancing individuals' sense of personal efficacy could counteract the effects of actual and felt powerlessness vis-à-vis an authority (Kelman & Hamilton, 1989: 323).

This argument also resonates when thinking about digital authority. It was exactly this sense of self-efficacy that made Stanislav Petrov say no to the digital system telling him a missile strike was about to hit the U.S.S.R, just like confidence in their own capacities formed an incentive for the Kentucky judges who refused to rely on the risk assessment technology. An important step towards enhancing personal efficacy in human-technology interactions is to open the block-box and teach users about the workings, limits, and potential failures of digital technologies. Education and training are one way to do so.

Handbooks, presentations, and discussion sessions are examples of methods to teach individuals about the possible risks and weaknesses of digital systems. Another way to make people aware of the possibility of digital failure is to have them *experience* it first-hand. This idea originates from what is called inoculation theory (McGuire, 1961). In medicine, inoculation refers to exposing someone to a weakened form of a virus (a vaccine) that does not overwhelm the body's immune system but rather will trigger a response (the production of antibodies) that will protect against the actual disease. As described by Van der Linden and colleagues (2017), a similar logic is followed in the social-psychological theory of attitudinal inoculation: a threat is introduced by forewarning people that they may be confronted with incorrect information. Subsequently, "one or more (weakened) examples of that information are presented and directly refuted in a process called *refutational pre-emption* or *prebunking*" (Van der Linden et al, 2017:2). By pre-emptively demonstrating false claims and refuting potential counterarguments resistance is conferred. Educational methods in which individuals experience first-handedly that sometimes technologies can be wrong and that it is not just others who could be fooled by illegitimate digital authorities, can stimulate critical judgement and the potential for user resistance (see also Sagarin et al, 2002).

Familiarising people with the (in)capabilities of a technology allows them to recognise those moments in which their special attention is warranted and provides them with potential solutions for a given problem but also with *confidence* needed to act decisively. When working with a digital authority, training could leave individuals feeling better qualified to evaluate digital communications – to evaluate, question and if necessary, challenge digital output, for example by seeking support from other colleagues or initiating further investigation. As such, technology education potentially enhances personal *efficacy*.

Hence, education and training – at least to a certain extent – aims to reduce the distance between individuals and digital authority, stripping it of its mystery by revealing

what it can and cannot do. When closer to the technology, individuals are less likely to be overwhelmed by authoritative digital communications. They are reminded that, like human beings, digital authorities can be wrong. It is important to note that technology education can be general as well as specific, and ideally people living in a *digitope* are subjected to both. That is, professionals can learn about a specific digital technology and its potential benefits and risks. They are made more familiar with the technology and are provided tools for interacting with it responsibly. At the same time, education can also be dispersed more widely, to all segments of our digital society, to counter the tendency in all of us to sometimes rely on digital technologies blindly. Ideally, children in primary and secondary education are enrolled in programmes directed to making people more familiar with our digital world and the authority of things.[129] They help them to develop tools for getting by in a digital society and stimulate a healthy form of *digital skepticism*. Such digital skepticism revolves around the idea that people learn that there is always the possibility that digital output sometimes is unverified information, and that the utilisation of other available information sources is sometimes fruitful (see also Sagarin & Wood, 2007: 321). Future research could be directed towards the development of methods for stimulating such healthy digital skepticism without it becoming digital cynicism. After all, the moments we need digital technologies to make good decisions will only continue to increase.

In addition to education, ensuring enough decision-making time is also an important factor to enhance a personal sense of efficacy. In chapter 4, the notion of epistemic niche was discussed. Here, individuals are severely constrained in acquiring relevant insights from other information sources. Factors facilitating such a niche are, among others, substantial workload, and limited time availability. High workloads and little time limit the opportunity for individuals to evaluate digital output and could contribute to an experienced inequality in the relationship with the technology- feeling that they have no reasonable choice other than to comply with the technology's communications. Some studies have also found that compliance with decision aids becomes more likely when individuals' decision time is reduced (Rice & Keller, 2009; Hoff & Bashir, 2015: 416). Conversely, lower workloads could in some environments open up the epistemic niche and potentially leave individuals better equipped to come up with their own judgements and more confident to say no to a digital technology. Surely, not every decision can be delayed, and time pressure will always be part of the work of many professionals who often must decide on the spot – the parking officials from paragraph 4.4 are a case in point. In these, and all other, decision-making contexts it is important to structurally organise feedback and analyse results *ex post*. This allows those involved to reflect and learn from their own and others' decisions that and these insights could help improve future decision-making (see also section 7.7.3 below).

7.6.2 Empowering human beings

Sometimes individuals' critical judgement is impaired by the belief that they do not have a *right* to challenge authority – they feel that they are in no position to think and act for themselves. Kelman and Hamilton (1989: 322) use the term *empowerment* to refer to the process of securing the positions and powers of individuals that gives them the opportunity and the right to make their own decisions. Empowering individuals is about giving them the power to do something and stimulating them to develop an independent perspective.

Organisations and designers play an important role here. They can empower individuals by making explicit that information systems are relevant sources of information, but that there is no general obligation to follow them. An example is a facial recognition system where a possible identification - a match between a recorded image and photos in a database - is accompanied with capital letters saying:

> "This is not a positive identification. It is as an investigative lead only and it is not probable cause for arrest" (Hill, 2020).

Such messages emphasise professional roles and the agency that comes with it – that people have the right and power to act otherwise - when good reasons present themselves. In fact, organisations could even go further and make explicit it is not only the professionals' right, but their *duty* to think about what a digital system presents to them. Placing an emphasis on the responsibility of the individual – or what it means to be a "good professional" - could contribute to making people less susceptible to automatic deference to digital authority. It highlights that being a good professional does not equal blindly relying on a digital technology, just like it is not a sign of organisational disloyalty when someone seeks to critically evaluate a computer's output. While professionals who structurally and unquestionably act based on whatever a digital technology provides them might be attractive for some organisations from the point view of control (e.g., knowing users' responses and the ability to steer them) and efficiency (e.g., increased processing of cases), it does come with the risk of both individual, organisational, and societal harm. Therefore, when professionals are presented with doubtful output from a digital authority, the organisation can position itself to function as a "higher authority" and provide alternative interpretations of what a good professional is and must do. This way, the felt sense of obligation might become more directed at the organisation and its overall mission rather than to the technology itself.

Empowerment also involves providing sufficient social support for critical decision-making. Offering practical guidance, procedures and transparent structures that allow people to use their powers and challenge a technology's communications. Such structures could even be designed into the technologies – bringing us back to the notion of techno-regulation as discussed in paragraph 4.7. For example, people can be required to consult

a colleague before finalising a decision or must regularly check in with a supervisor. Implementing guidelines and mandatory moments of human evaluation into the technologies would not just remind individuals of their own powers, but also stimulate (or force) them to put these into practice when deemed necessary.

Kelman and Hamilton (1989: 322-3) explain that efficacy and empowerment are in a sense related. Empowerment can increase the skills and other resources needed for effective action and contribute to a sense of personal efficacy, while efficacy, in turn, can increase the resources needed for effective action and can increase someone's capacity to seize opportunities to use his or her powers and rights. Both empowerment and efficacy can also be linked, respectively, to the sources from which authorities (digital or human) typically derive their capacity to be consequential: the hierarchical position and coercion associated with it on the one hand, and knowledge and expertise on the other. Kelman and Hamilton argue that by means of empowerment and enhanced efficacy, individuals acquire some of the elements of authority themselves (1989: 323). Thinking about digital authority relations, both empowerment and enhanced efficacy potentially alter the relationship between human beings and technologies.

7.6.3 Intervention and compensation: handling a wrongful yes

Placing machines in the frontlines of (security) governance can have adverse effects. This can be the result of human decision-makers following a computer's output in a largely automatic and unthinking fashion and fail to recognise dubious claims even though they are faced with considerable evidence of this. But harm can also be an inevitable consequence of living and working in a *digitope* where digital authority is the result of technologies having access to a plethora of information that human beings remain barred from. Under these circumstances individual decision-makers sometimes have little rational alternative than to listen to the communications of digital technologies. Relying on digital output, then, is often the most reasonable thing one could do – even though there is no way of knowing it is right (as illustrated in paragraphs 4.4 and 4.5). And sometimes this will lead people astray and cause harm. But even though mistakes are bound to happen, this does not mean we must not seek to mitigate any negative consequences. This holds true for both individuals as well as for organisations and society at large. How do we get from saying yes, to admitting we have been wrong?

Let's start with the role of the individual's involved. While those who work with digital technologies sometimes have little choice than to accept inaccurate digital output, there is often a choice in how to do it. In many situations, professionals have at least some discretion in how to execute authoritative digital communications. As an illustration, consider some of the examples that have been discussed throughout this book – e.g. the mistaken arrests on the basis of a false facial recognition system's hit, the people who are refused to board a plane because their name is on an inaccurate digital no-fly list, police officers who shoot a

driver on the basis of a false digital claim that the vehicle was stolen, or the woman with the poodle whose car was aggressively boxed in on the highway after a computer had flagged her as potential drug trafficker. In these situations, the security officials were provided with information by a computer they could hardly ignore; they had no previous insights into the criminal backgrounds of these people and there was little or no way for them to validate the output. Yet, they were relatively free in deciding how to put that information into practice. When one can impossibly know about whether digital communications are correct, the path of action should be that of minimal potential harm in case the technology turns out to be incorrect. This means that professionals must always remain mindful about how to treat and approach those who have been flagged a security threat. With the possibility of information systems being wrong - and thus that people being stopped, searched, or arrested are innocent – the actions that follow from relying on digital authorities must always be executed as carefully and respectfully as possible. The subsequent chapter contains some additional reflections on the ethics of digital authority.

In line with the previous section, organisations and societies, in turn, must ensure there are sufficient guidelines and regulations that guide people in employing digital output – stimulating them to remain cautious when putting digital output into practice. And there are, at least, two additional responsibilities here. First, it is important to ensure that there are sufficient resources in place to discover and recover digital flaws as quickly as possible. There are many instances in which mistakes were discovered only very late or fixing things was not prioritised enough by organisations. The right responses were claimed to be difficult, for example from a technological - the case of the mistaken vehicle registration (NRC 2018) - or ownership (Nissenbaum, 1994) perspective. Either way, people ended up trapped in a bureaucratic web and unable to move forward (Widlak & Peeters, 2018). Or they remained unaware of the fact that harm was done in the first place – think of the digital errors of the Washington state prisons that were known for years before they were made known to the public.

In sum, even when the initial problems cannot be prevented, negative consequences must be limited through direct intervention and restoration of the inaccurate information in the technology and through compensation for those who became victims of such errors. This also makes that for our digital society, there must be institutions that help individuals who have been, in some way, wronged by digital authorities. The presence of independent oversight institutions and accountability mechanisms is indispensable to a just digital society.

7.7 CONCLUSION PART 3

Digital authority comes with clear limits. In the relationship with a technology, human expertise, system knowledge, institutionally ascribed authority or technological failures can hinder the experienced asymmetry that is endemic to authority. In situations, the absence of the right situational cues (e.g. design) or the disobedient responses of others can prevent people from defining something an authority situation. Sometimes there are also confrontations with multiple authority claimants, and only one gets recognised. Looking at authority's cultural layer, there is some evidence that in certain cultures people are generally less inclined to rely on digital technologies. Similarly, institutional norms, scandals, and incidents, too, could limit the construction of digital authority – especially in a context of modern reflexivity where digital authorities, too, can always be challenged and replaced. Lastly, digital authority is limited by technologies' ability to recognise, regulate, and utilse emotions to (re-)claim authority. It leaves digital authority (still) generally fragile.

Building on these limits, enhancing human efficacy and empowerment can mitigate the risks of digital authority. Moreover, individuals, organisations and societies limit negative consequences through effectively restoring inaccurate information and ensuring compensation for flawed outcomes. This is summarized in table 3 below.

Table 3. Limits of digital authority

	Relation	**Situation**	**Culture**
Characteristics	Experienced relational asymmetry	Defining an authority situation	Expectations of authority
	Consequential communications	Impressions	Conditioned values and beliefs
	Differentiated access to knowledge and force (persuasion and coercion)	Symbols (uniforms, titles, insignia, behaviors, places, etc.)	Dynamic and open to change
Digital Authority	Digital authority as asymmetrical human-technology relation	Digital technologies tapping into pre-existing authority scripts.	Narratives on authority and technology
	Digital communications to be believed or followed	Computers are social actors	Risk society and reflexivity
	Differentiated access to knowledge (epistemic niche/ opacity data and algorithms)	Situational authority cues (positions, titles, symbols, behaviors, etc.) from digital technologies	Contingency and uncertainty facilitating openness to digital authority
	Digital access to force (authorisation/ technoregulation)		
Limitation	Ascribed human authority	Absence of right authority cues	Cultural openness to innovation
	Expertise and system knowledge	Peers Subject salience	Incidents and scandals
	Technological failures and inconveniences	Conflicting (human/digital) authority claims	Reflexivity
			Organisational norms
	Social and emotional intelligence in digital technologies		
Mitigation	Efficacy (Training, education, workload, decision-making time, etc.).		
	Empowerment (Messages, practical guidance, procedures, etc.).		
	Intervention & compensation negative consequences		