This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No.: 607480

LARGE SCALE INFORMATION EXPLOITATION OF FORENSIC DATA

# Artificial Police Agents: Looking Awry

**LASIE Project**

*FP7 - SEC-2013.1.6-1 - Framework and tools for (semi-) automated exploitation of massive amounts of digital data for forensic purposes – Integration Project*

**Grant Agreement n°: 607480**

Start date of project: 1 May 2014

Duration: 42 months

Document. ref.: n/a

Peace Research Institute Oslo (PRIO)

Oslo, 18 November 2016

# Artificial Police Agents: Looking Awry

A lunch seminar by Prof. Dr Mireille Hildebrandt[1]

*A public report by Stine Bergersen (PRIO)*

## Introduction

## Background to the LASIE project

LASIE, or 'LArge Scale Information Exploitation of Forensic Data' (2014-2017), is a research project co-funded by the European Union (EU) under the 7th Framework Programme for Research and Development (FP7), which aims to design and implement an open and expandable framework that will significantly increase the efficiency of current investigation practices, by providing an automated initial analysis of the vast amounts of heterogeneous forensic data that analysts have to cope with.[2] LASIE will create a smart surveillance tool, a "gadget" that can be used by law enforcement authorities (LEA), and which integrates various surveillance sources: it will be able to be connected to multiple, heterogeneous databases: from Close-Circuit Television (CCTV) videos and publicly available data on social network sites to voice recordings, it will optically recognise handwritten text and will read through blogs. Eventually, its aim is to be widely used by the police post-crime to collect evidence for a given case. While LASIE has a noble aim (i.e. to efficiently and meaningfully use the data available to investigate a criminal offence already happened), it also opens up for abuse (e.g. to be used beyond purposes for which it has been designed).

PRIO, as a partner in LASIE, is responsible for making sure that the LASIE gadget is "societally ok", that it is ethically sound and legally compliant, most specifically when it comes to issues of privacy and personal data protection. Furthermore, PRIO advises on how to make sure that the evidence produced or compiled by the LASIE "gadget" is admissible in the court of law in Europe. In this regard, the challenge of ensuring due process of criminal law can be mentioned: how can we scrutinize LASIE as a machine, to prove that the evidence has been provided in a proper way, e.g. that it has not been tampered with? Thus, for example, LASIE will provide a log file, to demonstrate how the machine has processed and compiled the evidence.

[1] Mireille Hildebrandt is a Research Professor at the research group for Law, Science, Technology and Society (LSTS) at Vrije Universiteit Brussel (VUB). She holds a PhD in the philosophy of criminal law from Erasmus University Rotterdam, and a law degree from Leiden University. Hildebrandt also holds the (part time) chair of Smart Environments, Data Protection and the Rule of Law at the Institute for Computing and Information Sciences (iCIS) at Radboud University, Nijmegen, the Netherlands.

[2] Cf. www.lasie-project.eu and http://cordis.europa.eu/project/rcn/185486_en.html.

## Background to LASIE seminars at PRIO

Since 2016, PRIO – as a partner in LASIE – organises short seminars at its premises in Oslo, Norway to discuss with invited experts the most pressing societal issues arising from the LASIE research project. The adopted formula of such seminars involves an introduction to a given topic by chairperson (15 min.), a keynote by the invited expert (45 min.), up to two interventions from invited commentators (10 min. each), the response from the invited expert (10 min.) and, eventually, the discussion with the public (30 min.) The seminars are organised either during a breakfast or a lunch and last 2 hours. Each seminar is followed by a public report such as the present one.

The first seminar hosted Prof. Dr Roger Clarke (Australian National University and the University of New South Wales) on 29th August 2016 and analysed the notion of "security".[3]

## Background to the lunch seminar with Mireille Hildebrandt

A lunch seminar by Research Prof. Dr Mireille Hildebrandt, with the title "Artificial Police Agents: Looking Awry", was on 18th November 2016 arranged at PRIO premises in Oslo.

The seminar was advertised internally at PRIO, on the public websites of PRIO, as well as internally within the LASIE consortium with the following description:

> *We now live in a world saturated with artificial agency, even though it is difficult to identify the agents. Police intelligence is increasingly supported by big data analytics and machine learning, e.g. in the form of crime mapping and new types of criminal profiling. This generates new issues around opacity and transparency that are key to the safety and security of individual citizens. In this talk, Mireille Hildebrandt will discuss the implications for privacy, non-discrimination and the presumption of innocence as well as the methodological pitfalls that plague the employment of computational optimization. Though the latter term is less intuitive than 'machine learning', it may be time to acknowledge that 'learning' can also be a misleadning term that diverts attention from the invisible backend of technical systems that are employed. Such systems profile presort individuals which may lead to further scrutiny and monitoring. As this part of investigations does remain hidden and does not fall within the scope of the presumption of innocence, it reconfigures the institutional landscape of the Rule of Law. Hildebrandt's argument will be that we must learn to 'look awry' at the input, the output and the computational operations of artificial agency, scrutinizing their claims to objectivity and truth without demonizing their employment.*

The present report offers a succinct account of the presentation by Hildebrandt, as well as some views, questions and comments expressed following the presentation. This report, as well as being a summary of the event also of interest to the LASIE consortium partners not present, also holds the potential to stimulate further debates on the wider impacts of technology on the society or the use of new technology by law enforcement authorities or intelligence community, e.g. linked to both the practical and theoretical implementation of the LASIE framework.

[3] Cf. http://www.lasie-project.eu/wp-content/uploads/2015/05/LASIE_Clarke_seminar_FINAL_ok_cleanv2.pdf.

Participants were encouraged to familiarize themselves with Hildebrandt's recent chapter in The SAGE Handbook of Global Policing[4].

The views expressed in the following are solely those of the speakers, and does not reflect the views of the LASIE consortium as a whole, its individual partners nor the European Commission.


**Figure 1: Photo by PRIO**

## Opening of the seminar

Dariusz Kloza, the LASIE project leader at PRIO and the chair of the session, opened the seminar by welcoming the participants, and by giving a brief introduction to PRIO, and the kind of research conducted at the Institute. In short, PRIO conducts research on the various conditions for peaceful relations between states, groups and people. As a backdrop for the theme of the seminar, within the department "Dimensions of Security", where the research team[5] in LASIE is situated, the focus is on critical security research, where several concepts of "security" are at play. "Security" is understood and explored at PRIO both as national security e.g. at borders, within cities, as societal security, etc. That is to say that no predominant perspective on the concept is established. While there is no doubt that new technology, "big data" and criminal profiling continue to be important topics for discussion in many and various disciplines, having this seminar in the context of the LASIE project is especially relevant because of the take on global policing, e.g. highlighting the idea that predictive technologies might generate more uncertainty rather than more certainty.

[4] http://sk.sagepub.com/reference/the-sage-handbook-of-global-policing/i2880.xml.

[5] Dariusz Kloza, Rocco Bellanova, Stine Bergersen and Ida Rødningen.

Kloza introduced the guest speaker, Prof. Mireille Hildebrandt, as a prolific academic writer, who researches and writes on law and technology, especially on "smart environments" and on the architecture of modern law. She has been part of the so-called 'onli*f*e initiative' of DG Connect,[6] and has further developed the concept of "onli*f*e world" (as opposed to "online" or "offline" world), where this one letter difference makes a big transformation in how we perceive machines, algorithms, etc. She has explored this phenomenon in her recent book "Smart Technologies and the End(s) of the Law" (2015),[7] and at this seminar, the overarching topic for her presentation was data-driven machines and human society. Kloza highlighted how the topic is both timely and relevant, and although a more romantic or popular view on artificial intelligence, robots, etc. has been dominating in movies and literature in the last century, the emerging of gadgets such as Google's Allo and Apple's Siri as well as the concept of 'deep learning' become part of everyday life. While such devices are on the one hand impressive and useful for many individuals, on the other hand – as for example expressed by Edward Snowden via his Twitter account – these devices engage in massive data collection based on their need to learn. This makes the use of machines, androids, robots and so on, very ambivalent, and raises many legal, ethical and societal questions. These are issues that Hildebrandt is working on.

Kloza also introduced the two discussants, Dr. Kristoffer Lidèn, senior researcher at PRIO, and Dr. Mareile Kaufmann, post-doctoral researcher at the University of Oslo, before giving the floor to Hildebrandt.

## Presentation by Prof. Mireille Hildebrandt: Artificial Police Agents: Looking Awry

Hildebrandt provided the audience with a brief introduction to her background, and her key research topics. She informed that her presentation was tweaked from being more about policing to saying something about forensics, to fit better the focus of the LASIE project. She also referred to an invitation from the Dutch association of lawyers, to write a so-called Pre-advise on "homo digitalis" in relation to criminal law,[8] which ends with two propositions or recommendations that Hildebrandt promised to return to at the end of her presentation. A central idea is that if you want to provide legal protection, rules and rights, as expressed in writing in law, texts etc. may require translation into the very architecture and infrastructure of the technologies and systems that are currently being developed, e.g. by the police, or in the case of LASIE, by the forensic institutes and forensic researchers that are building, buying and investing into such systems.

With regards to current research on forensics and machine learning, a handbook by Jesus Mena from 2016 was used as an example to see how forensics and machine learning relate to each other.[9] The chapter on Digital Investigative maps and models refers to the need to have a strategy in place, especially with regards to data collection, before doing machine learning. A difference is drawn up in the book between passive (post-crime) and active (predictive) forensic investigations, and further, a long line of methodological approaches are presented, such as inductive forensics, deductive forensics, and a range of cybersecurity tools. The author also provides a definition of forensic science, in relation

[6] Luciano Floridi (ed.). "The Onli*f*e Manifesto. Being Human in a Hyperconnected World", Dordrecht: Springer 2014.

[7] Cf. http://www.e-elgar.com/shop/smart-technologies-and-the-end-s-of-law.

[8] Mireille Hildebrandt. "Data-gestuurde Intelligentie in het Strafrecht", Deventer: Handelingen Nederlandse Juristen-Vereniging, Vol. 146 (2016), pp. 139 – 240. Available at: http://works.bepress.com/mireille_hildebrandt/68.

[9] Mena, J., 2016. *Machine Learning Forensics for Law Enforcement, Security, and Intelligence*, CRC Press.

to machine learning, as: the prediction of criminal intent and/or action (thus placing the whole field within predictive forensics).

In preparation of the writing of the above-mentioned "Pre-advise", Hildebrandt looked into what types of data driven operations the police in the Netherlands currently employ: the mining of social media, which can very soon be extended to the Internet of Things, with smart energy grids (for power usage) as a good example. The smart grids collect data about the energy usage behaviours of energy consumers, to predict their usage; in the Netherlands they can opt out of having a so-called smart meter, but this will result in being profiled based on what computer scientists call their nearest neighbours (other consumers that share some data points, such as family size, type of house, neighbourhood). Further, the mining of location data is not about where you are today, but the prediction about where you might be tomorrow. While we are used to thinking that content analysis is more privacy-intrusive than meta-data analysis, this is not necessarily the case. In some cases, the meta-data can be far more invasive, and say far more about what kind of person you are, than content data, which might not be that interesting. The police are also increasingly interested in remote access of computing systems, to compensate for the fact that communication is encrypted (they will access the information before or after it is encrypted, on the device of the sender or receiver). Hildebrandt underlined the importance of distinguishing between forensic science, in terms of discovery and evidence post-crime, and predictive analytics. In this regard, "big data" opens up many possibilities for researching different kinds of data and to mine all sorts of different patterns in the data that may have predictive value.

As a way of introducing the concept of machine learning, Hildebrandt presented the illustrative example of A-B testing. Machine learning can best be explained in reference to all websites that we experience as running smoothly and providing relevant information, as these are most probably continuously being A-B tested. This means that a provider has a website – version A – to which one small change is made, e.g. in layout, font, colours etc., - this is version B. The audience is split, and half the visitors are directed towards version A, the other half to version B. Software calculates which visitors' machine readable behaviour is most advantageous for the website owner, leading to a preferred version A or B. This preferred version then becomes the new version A, and can go through the same kind of testing. This process will be iterated to keep the site up to date, attractive and lucrative.

So, what is machine learning? When it comes to the very concept of machine learning, we can give the example of building software to grade papers, a need that arose after the introduction of some very popular online courses, resulting in a too big to handle amount of student papers to grade manually. How did they do this? The teacher grades the first 100 papers manually, then feeds both the papers and the grade to the machine, 'telling it' that these papers are the input, and that these grades are the correct output. Usually the text in the papers is labelled, to ease the task of the machine, which is to figure out the mathematical function that best relates the input to output, thus coming up with a statistical relation between the two (basically a hypothesis). Next, the teacher grades the next 200 papers, and 'asks' the machine to do the same thing. The teacher then compares his or her own grades to those of the machine, 'telling' it where it went wrong (this is an example of reinforcement learning). This whole process is continued – grading papers, feeding them into the machine, checking the output and adjusting – until the teachers observes that the machine gets it right. Furthermore, machines have the advantage of not getting angry, tired or frustrated, when a teacher might eventually be. On the contrary, the machine gets better, the larger the amount of papers is. However, teachers will still have to check the highest and lowest degrees, because e.g. the very best papers can be given the lowest

score, since those students might think in a completely different way, and thus be so brilliant and creative in their answer that the machine cannot calculate that. This type of machine learning is called natural language programming; it can be related to the fact that machines are getting better at reading legal texts, mining argumentation and predicting the outcome of court cases, threatening to take over parts of the work of legal services.

However, what is very important to note here, is that there will always be a training set (the data that form the input), and that data scientists will have to develop a set of hypotheses (mathematical functions) to enable the machine to relate the input to the output. The 'real' statistical relationship between input and output is called the *target function*, which is unknown – only by iterant testing of various hypotheses can one try to get close to the target function. The point is that the developer of the software needs the machine to predict the outcome for not only of the training set, but also for future sets. In fact, the target function cannot be known because of the fact that future data are not known. Hildebrandt referred to a computer scientist, David Wolpert, who developed in 1990s the "no free lunch" theorem. This is mathematical proof that since we never have access to future data, we can never be sure that a hypothesis function is the correct (the target) function. Thus, the optimization process may work on a particular training set for a particular output, but one can never be sure that it will work for another training set, if this concerns future data. As a consequence, the choice of the training set and the choice of the mathematical functions that you are testing, involve trade-offs and a kind of politics.

For example, the choice of a training set impacts the results that you get, e.g. if you use datasets that can be described as "low hanging fruit" (data that happens to be available or cheap, but not necessarily relevant), or if the datasets are very large (which can lead to detection of spurious correlations). In both cases the predictions are wrong, but may nevertheless influence people, which make it difficult to test whether or not they were wrong. Trade-offs are at stake between accuracy and predictive force (if the hypothesis function accurately describes the training data in great detail it might not generalize to the test set). Finding the right kind of hypothesis for your dataset is crucial for the quality of your predictions, taking into account that the predictions are in a sense always fragile, and needs to be tested iterantly. So, if you train your algorithms on a particular dataset, you might be getting incorrect results because the hypothesis works for the training set but not for the test set. If you are looking e.g. to predict homicide, you have to decide which data you think is important, you have to pick a set of data points that you are going to collect, and there is the possibility that there might be relevant data that you did not think to include in the training set. And if relevant data is not part of your training set, then this data will not be part of your output.

In short, the choice of training set co-determines the output. Furthermore, if the purpose of forensic sciences is to target *people*, e.g. as potential suspects, then this risk of wrong, incomplete, or false datasets is not merely a methodological problem, but a real problem that will have an impact on people's lives. From the perspective of data science, machine learning is based on detection of a relevant bias in a data set, and the bias is productive. If you cannot detect any bias, the data set would be random. In fact, the more data is put together, the more patterns (bias) will be found. The challenge, however, is to find the correct and relevant bias that can work for you in real life. The whole process of gathering data, processing them, building a hypothesis space and testing which hypothesis 'works' requires substantial investment, and considering the trade-offs and uncertainties one wonders whether this is all worth it.

Critics towards these systems often say that they should not have any bias, but central philosophical strains of thought claim that biases are important tools for us to make sense of the world. It is

important for humans to have frames of reference to discriminate between different situations, people, threats and opportunities, and to put these frames to the test. The question that we should ask is how the methodological issue of bias in machine learning relates to the ethical and legal understanding of bias and prohibited discrimination. What does it mean that these systems are always biased, and in a sense always have to be biased? What happens if the system is tested and a bias against e.g. Muslims is revealed? Then, the bias must be taken out, since it is prohibited to make decisions based on that factor. But many other types of bias may be acted upon, based on the operations of machine learning, that are not prohibited and that consist of complex mathematical calculations. Can we get our finger behind this and contest such bias? On what grounds?

It is important for the discussion on what machine learning *does* to first see that this technology has an enormous impact. It presents the police etc. with a specific – new – type of choice architecture (a term taken from nudge theory that wants to present people with specific options, that 'help' them to make better decisions). Machine learning has an impact because it reduces people's behavior to computable facts, and predicts their potentially criminal behaviour based on such 'facts', resulting in the police acting upon them. As we have seen such predictions are fragile and depend on all kinds of trade-offs. One of the dangers of predictive policing and forensics is to assume that data is the same as facts. A data is a translation, a representation or a trace of what it refers to, and the patterns found in the data are not necessarily to be found in 'reality' (whatever that is).

Technology is neither good nor bad, but never neutral. The first thing we must do is to look at the normativity which a specific technology generates. Normativity is not the same as morality, but refers to what behaviors a certain technology enforces, induces, precludes or inhibits. It refers to the impact of a technology on our behavior *patterns*. Once that normative impact is established, you can ask if that is good or bad (morality or ethics). Furthermore, it becomes a question of politics. What does data driven policing of data driven forensics afford? What does it rule out, what does it encourage, etc.? How does this affect checks and balances, between the police on the street (what criminologists might call street level work) and managerial policing? What does this imply for forensic researches? To what extent do they lose a certain discretion?

This brings Hildebrandt to the final point of the presentation, where she underlined that there are scholars who think that the police should not have any discretion, that they should just execute the law, with no room for own decisions (*legalism*). Hildebrandt disagrees with this position, and highlights the work of legal philosopher Dworkin, stating that actually the rule of law starts when policemen (or civil servants) act on the discretion to execute the rule of law: "discretion is not the absence of principles or rules; rather it is the space between them". If you take this flexibility away from the police, they will become predictable, and to the extent that this were to be possible they could be replaced robots or by software systems. However, this would imply a rigid, legalistic system, assuming that the future is already out there and that rules require no interpretation. This is a legalistic understanding of what the law is. If you allow forensic scientists to practice with some discretion, their expertise will grow while exercising that discretion. Indeed, their ability to raise important questions in a court of law as expert witnesses is very important in order to sustain and renew the integrity of forensic science. If you assume this, then the question we should ask, is how we can develop these sorts of systems in a way that enhances this integrity instead of "squeezing it"?

Finally, Hildebrandt shared the two propositions that she defended at the annual meeting of the Netherlands Lawyers Association, as mentioned in the introduction. These are based on the fact that, in the context of justice and police, the employment of predictive technologies results in iterant mining of massive amounts of personal data, leading to screening and monitoring of potential suspects – long

before any criminal offense has surfaced. This seems to violate the presumption of innocence, even if legally speaking this may not be the case. Thus, the first proposition was that in the new code of criminal procedure, rules must be imposed about the design and the default settings of these types of data driven systems, making sure they do not violate the rights and freedoms of individual citizens beyond what is necessary, foreseeable and proportional. The second proposition was that a watchdog must be established, that is tasked with the verification of the design, the default settings and the actual operations of these types of data driven systems, notably when employed for secret investigations. This means asking the question: to what extent do these systems enable the violation of rules, and how is such violation prevented (e.g. by means of secure logging and requiring the use of open source software)?

## Interventions from discussants

After the presentation by Hildebrandt, two invited discussants shared their comments and questions.

Mareile Kaufmann opened by complimenting the presentation for also being a great introduction to the field of machine learning. Kaufmann commented that she would reply from the perspective of her research on "deviance and the digital", and via three main points. The first point has to do with the assumption of "online world", and the distinctions between perception, agency and intelligence. Furthermore, Kaufmann raised the question about what is the role for humans in this. Humans *do* have a crucial role and an active part in writing the algorithms and correcting them, in creating the data and the systems, and thus in the politics that these systems and things create. In terms of biases, and *productive* biases, Kaufmann highlighted the self-fulfilling prophecy that occurs when you start policing on the basis of "wrong biases", so how e.g. these biases impact humans back should be discussed a lot more. Kaufmann's second point related to the reasoning in policing and law, and she asked what the ethical implications of applying this kind of correlational, aggregated, and patterned reasoning to individuals are. Her third point related to the writing and translation of rules into the system architecture (again, a point made in the distributed paper, but not in Hildebrandt's presentation), and concretely the question of what digital data (in Kaufmann's case relating to deviance) does to our lives? The answer is often that it is quantifiable, but what is won and lost with quantification?

Kristoffer Lidèn started by recommending the recent book by Hildebrandt as a great way into the field of machine learning and the wider field in general. Lidèn's intervention brought the topics back to basics, and focused on the politics involved. His first point, reflecting both recent developments and in preparation of future developments, refers to the long-term effects of small steps taken now (such as the introduction of GPS in mobile phones). Often, in EU security research projects, we (here: PRIO) are asked to evaluate the ethics of a certain system or device, and knowing that there are long-term effects associated with function creep, it can be challenging to make evaluations beyond the current context. In other words, Lidèn reminded us that although the implementation of a certain system or device might be acceptable as per the current rules and regulation, and also in the proximate societal context, it is almost impossible to ensure that this continues to be the case in future contexts with other constellations of variables and factors.

Furthermore, his first question was about the realism of implementing the kind of systems that Hildebrandt is describing. This question is indeed relevant for the LASIE project, because it is about the creation (and possibly implementation) of such a system, but the issue of theory versus practice is relevant also with regards to many other research projects that PRIO is or has been involved with.

Lidèn's second point was about the importance of the wider societal context that these systems are introduced within, a particular strength in Hildebrandt's research, according to him (such as the notion of "onlife world"). The point here is that the development and use of e.g. systems for predictive policing based on machine learning takes place in a societal context and environment, where you are likely to face a similar dynamics as with hacking and counter-hacking, that different groups (e.g. state versus criminals) learn from each other and seek to overdo each other. Thus, one could ask what you gain by starting such a race. One answer could be that it is not an option for the police to sit still and wait while this precedes in the general society, but there are nonetheless issues relating to legal regulations (e.g. how to avoid technologies created to do "good" can be picked up by actors with malign intentions) that should be considered. We might have a situation where both people and machines will learn about the police and police methods, and seek to manipulate them eventually (sort of a double-hermeneutics), leaving the system more vulnerable because malign actors might exploit e.g. the biases written into the systems. Although having biases can certainly be good, they are hard to control. The lack of neutrality is also an important issue.

## Discussion

Following these two interventions was a session where Hildebrandt responded to some of the comments and questions, and some questions from the audience were given. A very brief summary of this makes up the final section of this report.

After the two interventions, Hildebrandt provided some comments and reflections, e.g. with regards to the notion of "agency" (used on a high level of abstraction) highlighting that we as humans need to find a way of interacting with systems that, even though they are mindless, have agency, and influence us. We use them, they use our data, and we should now move from usage to interaction. Another issue raised by Hildebrandt was the over-production of data and patterns ("data and pattern obesitas"), resulting e.g. in that machines can find and interpret patterns that do not really mean anything. The current over-production of data means that it is important to discuss what data is really relevant and necessary to achieve the purpose of whoever is processing the data, also because the less data you collect the less security you need to install. Here we see that methodological integrity and data protection align.

Another point, summing up the comments to several points in the two interventions, related to the issue of legal protection by design. This concept implies that you design the system architecture based on (1) what the democratic legislator has decided, (2) enable resistance and (3) allows contestation in a court of law. It is based on the wish to articulate fundamental rights such as the presumption of innocence into the technological systems that threaten these rights, instead of trying to nudge people into compliance.

Finally, some questions and comments were received from the audience as well.

Due to the interest in the discussion, the seminar went a little bit over the scheduled time, but was finally rounded off by Kloza, thanking everyone for their interest and participation.