Intelligence Network & Secure Platform for Evidence Correlation and Transfer

D8.7 Privacy and Ethics-by-design in the INSPECTr platform

Document Summary Information

Grant Agreement No	833276	**Acronym**	INSPECTr
Full Title	**I**ntelligence **N**etwork & **S**ecure **P**latform for **E**vidence **C**orrelation and **Tr**ansfer		
Start Date	01/09/2019	**Duration**	42 months
Project URL	https://www.inspectr-project.eu		
Deliverable	D8.7 Privacy and Ethics-by-design in the INSPECTr platform		
Work Package	WP 8		
Contractual due date	31/08/2021	**Actual submission date**	31/08/2021
Nature	R	**Dissemination Level**	PU
Lead Beneficiary	TRI		
Responsible Author	Dr. Joshua Hughes		
Contributions from	Dr. David Barnard Wills		

Revision history (including peer reviewing & quality control)

Version	Issue Date	% Complete	Changes	Contributor(s)
v.0.1	01.06.21	0	Initial Deliverable Structure	Joshua Hughes
v.0.2	12.07.21		Substantive writing	Joshua Hughes
v.0.3	16.07.21		Review, privacy-by-design in LEA technology, section 4.3 added.	David Barnard-Wills, David Wright
v0.4	08.08.21		Update from reviewer comments	Joshua Hughes
v0.5	12.08.21		Updates	Joshua Hughes
V1.0	25.08.21	100	Update following UCD DPO review	Joshua Hughes

Disclaimer

Copyright message

Table of Contents

List of Tables

Glossary of terms and abbreviations used

Abbreviation / Term	Description
CCI	UCD Centre for Cybersecurity and Cybercrime Investigation
DMP	Data Management Plan
DPA	Data Protection Authority
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
EAB	Ethics Advisory Board
EC	European Commission
EDPS	European Data Protection Supervisor
GDPR	General Data Protection Regulation
LEA	Law Enforcement Authority
LED	Law Enforcement Directive
LIA	Legitimate Interests Assessment
LSG	Law Enforcement Authority Steering Group
POPD	Processing of Personal Data
TRI	Trilateral Research
UCD	University College Dublin
WP	Work Package

1 Introduction

The aim of this deliverable is to introduce the concepts of Ethics-by Design and Privacy-by-Design to the INSPECTr project, to explain how these design approaches are incorporated into the INSPECTr project, and to show some of the design solutions that have been developed using these approaches so far in the project.

1.1 Mapping INSPECTr Outputs

The purpose of this section is to map INSPECTr Grant Agreement commitments, both within the formal deliverable and task description, against the project's respective outputs and work performed.

Table 1: Adherence to INSPECTr GA Deliverable & Tasks Descriptions

INSPECTr GA Component Title	**INSPECTr GA Component Outline**	**Respective Document Chapter(s)**	**Justification**
DELIVERABLE			
D8.7 Privacy and Ethics-by-design in the INSPECTr platform	*Report describing the privacy and ethics-by design measures and approaches as included in the INSPECTr platform and tools*	*Sections 2, 3 and 4.*	*Sections 2 and 3 discuss the nature of Privacy and Ethics-by-Design in INSPECTr, respectively.* *Section 4 documents some of the design solutions developed using Privacy and Ethics-by-Design approaches.*
TASKS			
T8.3 *Privacy-by-design and Ethics by design for INSPECTr tools and platforms*	*Provide Privacy-by-Design and Ethics-by-Design support to infrastructure and analysis tool development (primarily in WP3 to WP5) as needed, in an ongoing, responsive and agile basis. This is an iterative process following the known best practices, and parallel case studies that deal with potential privacy, ethical and social impacts. Identify potential impacts at the different levels of the design process and mitigate negative impacts. This will include:* *• Responding to legal or ethical queries as they*	*Sections 2, 3 and 4.*	*Sections 2 and 3 document the concepts and processes of Privacy-by-Design and Ethics-by-Design.* *Section 4 documents how these processes are implemented in the INSPECTr project, and gives examples of how this process has resulted in changes to the INSPECTr tools, platform, and project.*

	emerge from the design and development process • *Monitoring the technology design and development processes to identify any emergent privacy or ethics issues and collaborating to produce design solutions.* • *Acting as an internal stakeholder for privacy and ethics related issues in the project.* • *Identifying relevant resources of designers and developers (e.g. privacy-protecting design patterns).*		

1.2 Deliverable Overview and Report Structure

This deliverable has three main sections.

Section 2 explains the concept of Ethics-by-Design and how it is applied in INSPECTr. Section 3 does the same for Privacy-by-Design.

Section 4 sets out how this process works in INSPECTr and displays the main ethical and privacy issues that have been discussed and resulted in design solutions during the project so far.

2 Privacy by design

2.1 Historical development

Historically, privacy concerns and technology design have been considered as separate issues: technologies are research, developed, and built by technology developers, and the use of these technologies to purposely invade privacy, or fail to protect privacy, have been the concern of lawyers and privacy advocates.[1] This separation has led to powerful technologies being placed into the hands of people and organisations who do not seen privacy as a main concern. This includes, for example, large technology companies who have the ability to track how we use their products and interact online through advertising infrastructures, but also organisations who could see privacy as a barrier to their work, such as intelligence agencies, and, in some cases, law enforcement. Where privacy is considered, it is often seen in a trade-off with other values such as security, accuracy, or commercial profit.

Privacy is, however, a human right[2] as well as an ethical and social concept closely tied to our human dignity,[3] and so it must be respected and protected. The concept of 'Privacy-by-Design' imbues technology development with privacy concerns. This builds on the several decades of research on privacy enhancing, or privacy respecting, technologies.[4]

Privacy-by-Design was first used by then Information and Privacy Commissioner of Ontario, Dr. Ann Cavoukin who suggested 7 foundational principles:

1. **Proactive not Reactive**; *Preventative not Remedial*. Anticipate, identify and prevent privacy invasive events before they occur. This principle aligns with the strong requirement from privacy impact assessments that privacy (and ethical) issues be considered early in the design process, including at the stage where initial ideas and objectives are being considered.[5]
2. **Privacy as the Default Setting**. Build in the maximum degree of privacy into the default settings for any system or business practice. Doing so will keep a user's privacy intact, even if they choose to do nothing.
3. **Privacy Embedded into Design**. Embed privacy settings into the design and architecture of information technology systems and business practices instead of implementing them after the fact as an add-on.
4. **Full Functionality** — *Positive-Sum, not Zero-Sum*. Accommodate all legitimate interests and objectives in a positive-sum manner to create a balance between privacy and security because it is possible to have both.

---

[1] European Data Protection Supervisor, Opinion 05/2018 Preliminary Opinion on Privacy by Design, paras.14-15. Available at: https://edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf

[2] See, Art.8, Convention for the Protection of Human Rights and Fundamental Freedoms (adopted 4 November 1950, entered into force 3 September 1953) 213 UNTS 221; Art.7, Charter of Fundamental Rights of the European Union, OJ C364/1, 18 December 2000.

[3] Floridi, L., "On Human Dignity as a Foundation for the Right to Privacy", *Philosophy & Technology*, Vol.29, 2016, pp.307-312.

[4] See, for example, Chaum, D., "Security without Identification: Card Computers to make Big Brother Obsolete" Communications of the ACM, vol. 28 no. 10, October 1985 pp. 1030-1044

[5] Wright, D., and de Hert, P. (eds.), *Privacy Impact Assessment*, Springer, Dordrecht, 2012.

5. **End-to-End Security** — *Full Lifecycle Protection*. Embed strong security measures to the complete lifecycle of data to ensure secure management of the information from beginning to end.
6. **Visibility and Transparency** — *Keep it Open*. Assure stakeholders that privacy standards are open, transparent and subject to independent verification.
7. **Respect for User Privacy** — *Keep it User-Centric*. Protect the interests of users by offering strong privacy defaults, appropriate notice, and empowering user-friendly options.[6]

Following these principles incorporates privacy concerns as a key element of the design process to ensure that technologies are developed and built in such a way that the privacy of end-users and people who might be affected by the technology is respected as far as practicable. Following these principles can facilitate end-user trust in a product or service where people would otherwise be concerned about a loss of privacy.

2.2 Legal manifestation - Data protection by design and default

The concept of Privacy-by-Design has been supported by the Article 29 Working Party[7] (now the European Data Protection Board), and the European Data Protection Supervisor.[8] However, as an ethical and societal concept Privacy-by-Design lacks the reinforcement of being a binding legal requirement. Thus, the concept has progressed to incorporate 'data protection by design' and 'data protection by default', which can be seen as the legal obligations within the wider Privacy-by-Design concept.[9] They are included in Article 25of the GDPR, which provides:

> '*1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.*
>
> *2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular,*

[6] Information and Privacy Commissioner of Ontario, "Privacy by Design", January 2018. Available at: https://www.ipc.on.ca/wp-content/uploads/2018/01/pbd.pdf

[7] Cited in European Data Protection Supervisor, Opinion 05/2018 Preliminary Opinion on Privacy by Design, para.20. Available at: https://edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf

[8] European Data Protection Supervisor, Opinion of the European Data Protection Supervisor on Promoting Trust in the Information Society by Fostering Data Protection and Privacy, 2010. Available at: https://edps.europa.eu/sites/edp/files/publication/10-03-19_trust_information_society_en.pdf

[9] European Data Protection Supervisor, Opinion 05/2018 Preliminary Opinion on Privacy by Design, para.4. Available at: https://edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf

such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons. …'

One reason for including this in the GDPR as a legal obligation is the belief that the way in which the architectures of data processing systems are designed has great regulatory potential in order to protect people's personal data.[10] By placing a specific type of Privacy-by-Design in the GDPR, this places the obligation for compliance onto data controllers. [11]

This provision is, however, not specifically aimed at the designers of technologies and goods. Rather, it is aimed at all data controllers to design their processes and systems to facilitate compliance with data protection legislation. Yet, in INSPECTr the focus is on the technical partners as researchers and developers of data processing tools. Recital 78 of the GDPR states that:

'*When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations.*'

As such, INSPECTr partners should build their tools for the proposed platform in such a way as to *facilitate* end-users complying with their data protection obligations.

Adoption of the INSPECTr tools will not be possible by end-users if the tools do not facilitate LEA's compliance with their legal and operational obligations – including making it straightforward to evidence that they are doing so. Another imperative is important to recognise, the data protection by design and default requirement under the GDPR is mirrored in Article 20 of the Law Enforcement Directive (LED).[12] Therefore, the end-users of the INSPECTr tools will be obligated to choose systems, and comply with the functionalities of those systems, that allow them to abide by their data protection obligations under their Member State law implementing the LED.

2.3 Privacy-by-design strategies and techniques.

In parallel with the legal definition, more recent work has attempted to make these principles more practical and provide guidance for systems designers to better apply privacy by design. Projects have attempted to curate examples of privacy by design patterns that could be re-used and applied.[13] One of the more developed elements of this is eight key privacy design strategies for software

[10] Bygrave, Lee A. "Article 25", in Christopher Kuner, Lee A. Bygrave, and Christopher Docksey (eds.), "The EU General Data Protection Regulation (GDPR): A Commentary", Oxford, OUP. 2020, pp.573

[11] The GDPR is really a "personal data protection" regime rather than a "privacy" regime per se.

[12] European Parliament and Council, Directive (EU) 2016/680 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, OJ L 119/89, Vol.59, 4 May 2016 (Law Enforcement Directive, hereafter: LED)

[13] See, for example, PrivacyPatterns, UC Berkeley, School of Information. Available at: https://privacypatterns.org/

development offered by ENISA, the European Cybersecurity agency,[14] and privacy engineering expert Hoepman.[15] They provide the following:

Data Orientated strategies:

1. **Minimise** – the amount of personal data collected and processed should be as little as possible. Practically, this can include implementing 'select before you collect' guidance, and the use of anonymisation or pseudonymisation.
2. **Hide** – where personal data, and the interrelationships between them, do not need to be seen, they should not be. Practically, this can include encryption, anonymisation, and pseudonymisation, but techniques to un-link related personal data are also relevant.
3. **Separate** – personal data should be compartmentalised and processed in a distributed fashion to prevent unnecessary correlations being drawn. There are no practical concepts for this, but data should not be processed centrally and should be split into different database or unit where possible.
4. **Aggregate/Abstract** – Personal data should be processed at the highest level of aggregation or abstraction possible. Practically, this could involve aggregating data or time or location, or anonymisation techniques such as k-anonymity or differential privacy.

Process orientated strategies

5. **Inform** – wherever data-subjects use a system, it should be clear what personal data is being processed and how. Practically, this can involve allowing end-users to select privacy preferences, notifying them of data breaches, and generally being transparent about the processing of personal data.
6. **Control** – data-subjects should be given control over how their data is used. Practically, this can involve user-centric data management and end-to-end encryption.
7. **Enforce** – legal obligations should be enforced, including the need for an accurate privacy policy. Practically, this can include access controls such as privacy rights management tools to license the availability of personal data to certain organisations for a certain amount of time.
8. **Demonstrate** – the data controller should be able to demonstrate compliance with their privacy policy which, in turn, complies with applicable data protection legislation. Practically, this can involve privacy management systems, logging, and auditing.

These design strategies are a good point of departure. However, it must be borne in mind that the INSPECTr tools are intended to be used by LEAs in investigation of criminal offences. Such investigations, by their nature, involve the discovery of previously confidential information in accordance with strict legal oversight. Consequently, whilst following the principles of Privacy-by-Design in INSPECTr, they need to be adapted to the context in which the tools will be used. Part of this comes through discussions with LEAs about what is needed in an operational context, and what appropriate limits might be. As such, whilst Privacy-by-Design for software developers is often about limiting the amount of personal data collected from end-users by the technologies, in the situation of INSPECTr it is implemented in such a way to allow legitimate violations of privacy to occur in lawful

[14] Danezis, George, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Métayer, Rodica Tirtea, Stefan Schiffner, *Privacy and Data Protection by Design*, ENISA, December 2014, pp.18-22; note that ENISA has a legally mandate role in implementing Union and Member State law on privacy and data protection. See Arts.5(5)(c) and 7(2), and Recital 41, Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2016 (Cybersecurity Act), OJ L 151/15, Vol.62, 7 June 2019.

[15] Hoepman, Jaap-Henk, *Privacy Design Strategies*, 2019.

investigations, with proper accountability and security, whilst limiting the privacy impact to that which is needed for investigation of specific suspects.

It is important to note that, with respect to LEAs, implementation of data protection by design and default cannot depend solely on economic considerations. As such, an insufficient budget for LEAs cannot be a limiting factor on its own for limited implementation of measures to protect the rights and freedoms of data subjects.[16] However, as the INSPECTr project intends to provide tools at no, or very little, cost, this should not be a concern for LEAs who will be able to access the tools created by the project. Further consideration of how such tools should be used following the project will be provided in D8.8 (Guide on privacy and ethics-by-design in law enforcement technology).

2.4 What space is there for Privacy-by-Design in law enforcement technology research?

As discussed above, privacy-by-design is an approach to technology and business model design that explicitly brings privacy into a design process at an early stage as an important value. It has the ambition of producing forms of technology that meet the goals of developers or clients and other parties. Privacy by design is best understood as a process, rather than an outcome (which distinguishes it from formal models such as zero-trust).

It is, however, an open question as to what extent there is space in the design, development and deployment of policing technologies, and particular in forensic technologies, for the inclusion of this form of privacy by design? Particular challenges are raised by the way privacy-by-design conceptualises non-zero-sum, "the user" and transparency.

Privacy-by-design as a concept has a rhetorical component – the aim of the concept of "zero-sum" is to show that exploitative models of data collection and use common in the private sector are not *necessary* to achieve their stated goals, but rather a product of poor design. Necessity is an important target because the concept of necessity plays a central role in determining the legality of many forms of data processing, particularly under legal regimes like the GDPR and Law enforcement Directive. If it is possible to show that an objective can be met with minimal personal data processing, then arguments that such processing is necessary lose force. Non-zero-sum argues that both commercial interests and individual privacy interests can be met, as long as our designs are sufficiently clever.

The challenge is that policing activities trump many privacy interests and may present a case of an inherent zero-sum situation. Police investigations want more evidence, and the criminals subject to them want less. A regular statement from LEA stakeholder is that their investigators may not know what information is important to an investigation until that investigation has concluded. Many of the strategies of data minimization, obfuscation or aggregation are seen through this lens as hobbling an investigator. The INSPECTr use cases from WP1 present several scenarios where connecting seemingly unrelated information allows a case to be better understood.

In many of the contexts envisaged by privacy-by-design, the "user" has choice about if they will use a service or not. The risk is that they are encouraged to use a service that will be harmful or disadvantageous for them in some way, or will exploit the data that they entrust it with. A classic example is a person signing up for a social media service that will use information about them to sell advertisers the ability to send highly targeted adds. Privacy-by-design in these contexts is in part a response to the problem of consent-based models, which use the legal consent of a user to justify exploitative activity, when it is well known that consumers do not read terms and conditions, and are very often not aware of how information about them will be used. The promise of privacy-by-design is that this tension can be resolved if the service is designed to be appropriately respectful of privacy.

[16] Recital 53, LED

Typically, police authorities do not need the consent of the data subject to process personal data in the pursuit of their legally mandated activities. Typically, policing activities are governed by different legislation (e.g., LED rather than the GDPR), and intrusions into private life are allowed that would not be allowed in a commercial context. Of course, police powers are not absolute, and fundamental rights and the courts act as a backstop to this, but it is quite distant from the model of privacy-by-design that originates from managing competing interests between commercial entities and their potential customers or users.

A related issue is that the consumer-facing principles of privacy-by-design also position the user and the data subject as essentially the same person, however this is not the case for law-enforcement technologies. The user for INSPECTr is a digital forensics analyst or investigator, whilst the data subject may be (inter alia) criminals, victims of crime or uninvolved people included in electronic evidence. There are very few anticipated impacts upon the privacy of platform users, but substantial privacy impacts for the various types of data subject. This is, however, not a hard principle to adapt and essentially requires interpreting as respect for the privacy of data subjects – which can largely piggy-back off compliance with data protection law, bolstered by considering how we can best further protect privacy of victims and bystanders even when an LEA might have legal authority to process their data.

Transparency should also be considered in terms of what is appropriate in the law enforcement situation. In the commercial context, transparency supposedly works by providing the user/consumer with information that they can use to decide if they trust their personal data to an organization or not. It also allows regulators to hold organisations to account. However, police authorities are under different social and legal requirements about transparency with relation to the data they process. For example, under the GDPR, transparency requirements obligate data controllers to inform data-subjects about the processing of their personal data.[17] However, in a law enforcement situation, an authority might only be required to publish information of a general nature about data analysis and only inform a data-subject where it is not '*liable to jeopardise the tasks for which those authorities are responsible.*'[18] In this sense, there is a connection between transparency and accountability as part of privacy-by-design in the law enforcement context. It is not immediate and prior transparency to a data subject to gain consent that is crucial, but rather accountability and attestability to lawful activities to a range of institutional actors with governance roles – senior officers, ethics boards,[19] data protection regulators, the judiciary, courts and, after investigations are over and information can be publicised, the public.

However, several privacy-by-design principles remain operationalizable in law enforcement technology (set out in the table in section 4.3 below). A proactive consideration of privacy, privacy as a default setting, embedding privacy into design can all be potentially implemented. End-to-end security is potentially even more important, given that such systems could be assumed to potentially under threat, as well as interacting with sensitive data. Additionally, data protection by design and default, as set out in the GDPR and mirrored in the LED is theoretically achievable, given the more limited mandate of developing systems that meet the requirements of data protection law.

[17] Arts.13 and 14, European Parliament and Council, Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, OJ L 119, Vol.59, 4 May 2016

[18] CJEU, C-511/18, C-512/18 and C-520/18, La Quadrature du Net and others [GC], 6 October 2020, para.191

[19] See, for example, West Midlands Police and Crime Commissioner, "Ethics Committee". Available at: https://www.westmidlands-pcc.gov.uk/ethics-committee/

3 Ethics by design

3.1 Historical development

Ethics-by-design has been largely influenced by both Privacy-by-Design and Value Sensitive Design as an approach to a design process. However, it must be noted that for some authors, the notion of 'Ethics-by-Design' for new technologies is about trying to incorporate ethical decision-making into algorithms.[20] Whether this is even possible is a technical, or perhaps philosophical, question. But, in this document, and in the INSPECTr project, the concept of Ethics-by-Design is used to refer to a design process in the same way as Privacy-by-Design is used.

As described above, Privacy-by-Design involves incorporating privacy concerns across the design process. Value sensitive design '*is a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process.*'[21]

In value sensitive design, technology development can be analysed in terms of what is 'good' or 'important', as determined by stakeholders, and alternative design options can be developed.[22]

It involves the following steps:

1. Start with a Value, Technology, or Context of Use
2. Identify Direct and Indirect Stakeholders
3. Identify Benefits and Harms for Each Stakeholder Group
4. Map Benefits and Harms onto Corresponding Values
5. Conduct a Conceptual Investigation of Key Values
6. Identify Potential Value Conflicts
7. Integrate Value Considerations into One's Organizational Structure

This process can lead to new design solutions that balance priorities from different values and mitigate harms. Importantly for this consideration, ethical values can be used as the main drivers of this type of design process. However, value sensitive design is predicated on analysing a technology that already exists and then developing alternative designs. Yet, if a harmful technology is developed then that would be a negative thing to happen, ethically speaking, whether or not that technology is actually used. It would be best for fulfilling ethical values that harmful technologies are not developed. Further, there is limited consideration of the purpose of the end-users, and the how the technology will contribute or take away from that purpose. Additionally, by focussing on what stakeholders determine is 'good', there is little consideration of how the development of the technology aligns with guiding principles of what is 'right'; something that is good for one person might not be the right thing to do if there are negative repercussions for others. As such, value sensitive design is a very good starting point, but additional features can be added to make the outcomes more ethical.

As discussed above, a key part of the Privacy-by-Design methodology is the regular implementation of privacy design strategies to make technology development more private, or more privacy-

[20] See, for example, Dignum, Virginia, Matteo Baldoni, Christina Baroglio et al., "Ethics by Design: necessity or curse?", in AIES '18 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, New Orleans, 2018, pp. 60-66. https://doi.org/10.1145/3278721.3278745, p. 60.

[21] Friedman, Batya, Peter H. Kahn, Jr., and Alan Borning, "Value Sensitive Design and Information Systems", in Ping Zhang and Dennis Galletta (eds.), M.E. Sharpe, London, England, 2006, p.349

[22] Friedman, Batya, Peter H. Kahn, Jr., and Alan Borning, "Value Sensitive Design and Information Systems", in Ping Zhang and Dennis Galletta (eds.), M.E. Sharpe, London, England, 2006, p.349, 366

respecting. With Ethics-by-Design, there is the implementation of ethical design requirements to try and make technology more ethical, or more aligned with ethical standards.

3.2 Ethical requirements and principles

A key issue in relation to ethical standards is whose standards should be chosen? A design team would gain limited benefit from following Nietzsche's argument that there is no rational foundation for morality, and there are no moral facts.[23] Generally, modern Western values seem to be chosen. This may, however, be because it is generally people in Western countries that are developing these guidelines due to their relative technological advancement.[24]

In the SHERPA project, partners analysed over 70 sets of potentially suitable ethical guidelines for Ethics-by-Design. This project was closely aligned with the EU's High-Level Expert group on Artificial Intelligence, and they share the same high-level requirements.[25]

Table 2: SHERPA High-Level Requirements

SHERPA Requirements and Sub-Requirements
1 Human agency, liberty and dignity: Positive liberty, negative liberty and human dignity
2 Technical robustness and safety: Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3 Privacy and data governance: Including respect for privacy, quality and integrity of data, access to data, data rights and ownership
4 Transparency: Including traceability, explainability and communication
5 Diversity, non-discrimination and fairness: Avoidance and reduction of bias, ensuring fairness and avoidance of discrimination, and inclusive stakeholder engagement
6 Individual, societal and environmental wellbeing: Sustainable and environmentally friendly SIS, individual well-being, social relationships and social cohesion, and democracy and strong institutions

[23] See Irwin, Thomas, *Ethics Through History*, Oxford, OUP, 2020, p.226

[24] See Jobin, Anna, Marcello Ienca, and Effy Vayena, "The global landscape of AI ethics guidelines", *Nature: Machine Intelligence*, Vol.1, September 2019, pp.389-399.

[25] See Philip Brey, Björn Lundgren, Kevin Macnish, and Mark Ryan, 'D3.2 Guidelines for the development and use of SIS', SHERPA project, 2019, p.1 (hereafter: 'SHERPA Guidelines'). Available at: https://dmu.figshare.com/articles/D3_2_Guidelines_for_the_development_and_the_use_of_SIS/11316833; High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, p.14. Available at: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf

7 Accountability:
auditability, minimisation and reporting of negative impact, internal and external governance frameworks, redress, and human oversight

These high-level requirements are useful and influential to the Ethics-by-Design process in INSPECTr. The SHERPA project went further than this to specify specific requirements for both the development and use of AI and big data systems. Again, these are inspirational to the approach in INSPECTr. However, due to the multi-party nature of the INSPECTr project, it would be difficult to apply these more granular requirements across different project partners who are carrying out different work simultaneously. As such, they are used as influences for the INSPECTr approach.

As noted above with respect to Privacy-by-Design, the tools being developed in the INSPECTr project are for a specific use by LEAs and so the design process needs to be specifically adapted. This is much the same with Ethics-by-Design. For example, transparency is a key element of the SHERPA requirements, but transparency with respect to LEA investigations needs to be tempered with a need to not alert criminals under investigation. As such, it can be beneficial to expressly consider Ethics-by-Design in terms of principles to be applied and then followed, rather than prescriptive requirements. This has the benefit that solutions can be tailored more specifically to the use technology under development, and in use. For example, The Ethics Centre provides the following principles:[26]

0. **Ought before can** – The fact that we can do something does not mean we should.
1. **Non-Instrumentalism** – Never design a technology in which people are merely a part of the machine.
2. **Self-Determination** – Maximise the freedom of those affected by your design.
3. **Responsibility** – Anticipate and design for all possible uses.
4. **Net Benefit** – Maximise good, minimise bad.
5. **Fairness** – Treat like cases in a like manner; different cases differently.
6. **Accessibility** – Design to include the most vulnerable user.
7. **Purpose** – Design with honest, clarity and fitness of purpose.

Bearing in mind that Ethics-by-Design is a state-of-the-art methodology, and that the INSPECTr project is not only incorporating ethical design approaches with AI ethics, but is doing so in the context of law enforcement, the Ethics-by-Design approach in INSPECTr is evolving as it progresses. This is important as the INSPECTr tools are also in development, and the Ethics-by-Design approach needs to be able to adjust in step with the technologies. In order to make those adjustments correctly, this requires consultation with stakeholders in the INSPECTr project to ensure that ethical design solutions that are suggested not only meet the requirements for ethical technology design, but also facilitate the proposed end-use of the INSPECTr tools.

[26] Beard, Matthew, and Simon Longstaff, *Ethical by Design*, The Ethics Centre, Sydney, Australia, 2018, p.59.

4 Ethics and Privacy-by-Design work in INSPECTr

4.1 How the Privacy and Ethics-by-Design process works in INSPECTr

As noted in the above two sections, Privacy and Ethics-by-design form part of the design process in INSPECTr which, due to being a project that researches tools for LEAs, means that a tailored approach is needed. Privacy and Ethics-by-Design approaches have been developed primarily for the development and regulation of technologies that are intended to be used by ordinary citizens, but the legitimate activities of LEAs during investigations can be intrusive and restrict people's freedoms.[27] Therefore, it is not possible to simply transplant an existing process into the INSPECTr project, as taking a level of privacy or ethical concern for ordinary civilian technologies and applying it to the exceptional situation of LEA technologies would likely render the technologies unfit for purpose: a data-analysis tool that could not accept data captured by intrusive surveillance, for example, would not be a useful result of the project.

Therefore, in terms of tailoring the approach to the INSPECTr tools, this requires that the purpose of the technologies is understood first and assessed to be an ethically acceptable goal which can be achieved in compliance with research ethics standards, and, if the technologies are intended to violate privacy when used after the project, that the legal and ethical limits of legitimate LEA investigations into people's private lives can be respected.

Next, an analysis needs to be made of how the tools work, and whether they can be improved in some way to make them: more ethically acceptable through mitigating potential harms; more privacy-respecting through reducing the ability of LEAs to use the tools in illegitimate intrusive ways, and increasing their ability to exercise better control and accountability around the use of these tools; less likely to create legal compliance issues through expected use, and less likely to facilitate unlawful use; more socially acceptable by minimising potential harm on a societal level. It is important to note that technology partners have not set out to create harms in the INSPECTr project, and any harms that might be created are inadvertent and might be unrecognised. A large part of the potential improvement of the INSPECTr tools involves highlighting harms and issues that could be created if the design of the tools is not altered and bringing these to the attention of developers. This allows them to bring their own problem-solving skills and domain knowledge into the Privacy and Ethics-by-Design approach.

Then, design solutions are developed in the context of the purpose for the tools. As noted above, a total focus on privacy would render tools for LEAs useless, and it is much the same for ethical, legal, and social concerns. Therefore, a balance needs to be struck so that these imperatives are included in the design considerations as much as possible, whilst also facilitating their use in the exceptional situation of LEA investigations. In some situations, this can result in changes to the technologies. But, in others it might not be possible to change the technology, either because it is at a state-of-the-art level and the required functionality needed to mitigate a harm might not exist or the technology would not work in a way other than it already does. Depending on the potential harms that could be created, an issue can be communicated to end-users so that it is taken into account during use; for example, potential algorithmic biases should be communicated to LEAs so that they are aware that they need to recognise that results of data analysis might not be completely objective. Or, a harm might be of such significance that, despite the best efforts of all partners in the project, it is not possible to mitigate issues to an acceptable degree; for example, a tool for detecting the emotion present in an image

[27] These would be legitimate and legal restrictions occurring even through appropriate use. For more information on scenarios around potential misuse see INSPECTr Deliverable 9.16: Risk assessment and measures to prevent misuse of research findings.

detection tool has been researched in the project and now will not be included in the final platform due to potentially significant issues associated with the capability of this technology.[28]

Examples of how issues about different tools have been highlighted and mitigated during the Ethics and Privacy-by-Design process are provided below. It should be noted that the process described above is evolving and additional steps might be added as the project progresses. Any required updates will be provided in D8.8 (Guide on privacy and ethics-by-design in law enforcement technology).

4.2 Implementation of requirements

Many of the design recommendations that have arisen from this process are captured in D8.5 (Ethical, Legal and Social requirements for the INSPECTr platform and tools). Several have been implemented across the Ethics and Privacy-by-Design process either in workshops, or during meetings with partners. But, as noted, this is an ongoing process where issues can be dealt with as they arise, rather than in a set sequence and so additional requirements can emerge that were not captured in the first group of requirements. A complete list of requirements and how they were fulfilled in INSPECTr will be included in D8.8 (Guide on privacy and ethics-by-design in law enforcement technology).

4.2.1 Workshops

At this stage, three ethics workshops have taken place in the project: use of publicly available (online) data; ethical AI, including discrimination and bias issues; gender and AI. During these events, many issues were raised and solutions found, although some issues required further consideration and research. Below is an explanation of some of the key issues that were dealt with in the INSPECTr ethics workshops. Note that some of the issues only apply to certain tools, or were discussed in terms of certain tools due to limited time, but solutions should be considered as generally applicable across all tools where possible.

4.2.1.1 Use of publicly available (online) data – January 2021

Web crawler

Collection and processing of open-source data from websites is an important part of contemporary LEA investigations. However, people can reveal significant amounts of personal data from posting on social media, online forums, and other websites. Collecting and processing such data can reveal intrusive insights into data-subjects. It was decided that it would not be appropriate to use the web scraping tools in the project for research. The tools is, therefore, restricted to use after the project by LEAs (an example of the Ought before can principle in section 3.2).

With respect to the technical capabilities of the tool, it is possible to collect a large amount of personal data from a website and this might be an unnecessarily large amount of personal data that is not needed in an investigation. It was discussed whether certain types of personal data could be excluded from collection. However, it was also noted that some LEA investigations grow larger as they progress and information not originally considered to be relevant might become important later on. Therefore, it was recommended that the web crawler should include a filtering capability that can hash personal data that is not needed, but could be revealed later in an investigation if needed (an example of a "hide" strategy from section 2.3). Or, the filter could be deactivated with permission of a senior officer where more complete web crawling might be required in an investigation (an example of an "enforce" strategy from section 2.3, and privacy as the default setting from section 2.1).

[28] See D4.6, Section 2.2 'Ethics', p.8

Availability of tools

As noted above the amount of personal data that people freely give away online can be significant. Where this data is analysed, highly-personal insights can be realised. Again, this might provide an intrusive level of information to an LEA investigator. Therefore, it was recommended that the INSPECTr tools are turned off by default and LEAs must determine which tools they need to use prior to data processing (an example of the privacy as the default setting from section 2.1). This might require authorisation from senior LEA officers about what tools are acceptable for each LEA to use in their country.

4.2.1.2 Ethical AI – February 2021

Bias

Due to the large amount of machine learning algorithms researched in the INSPECTr project, consideration needs to be given to the nature and types of data that are used to train the models. There is an ethical issue that where datasets contain data that is not representative of a particular population, then use of the algorithm that is trained with such data on that population is likely to have biased effects. It is important to note that poor data quality can also generate similar effects. For example, inadequate tagging and classification of objects in a computer vision tool could create similar issues.

For the national language processing (NLP) tools, it was discussed that error rates still need to be built into model outputs to help end-users understand them (an example of the Transparency requirement in section 3.2). Regarding computer vision models, it was determined that they could be fine-tuned to adjust for biases and then tested to determine if a model still functions poorly with certain groups. It was recommended that these tools are subjected to real world testing before being used. For the crime prediction tool, it was agreed that the tool could be adjusted for bias, but this would be difficult as some crime reporting does not reflect reality as some populations do not trust the police and so do not report crime as much as others; this needs to be considered in terms of the expectation that crime prediction tools are 'data agnostic' and the fact that data recordings is different across different LEAs (an example of the Diversity, non-discrimination, and fairness principles in section3.2).

Generally, it was recognised that some level of bias is an inevitable issue with (beyond) state-of-the-art machine learning technologies. There is not yet the technological capability to eliminate bias from machine learning models, and this will inevitably create disproportionate and biased impacts if/when INSPECTr tools are used after the project in the real world. Therefore, partners should conduct testing to determine the level of bias in their models, and whether it could be improved, and try to understand the impacts of remaining biases that can then be communicated to end-users so that they can take this into account when taking decisions.

Transparency and Explainability

Machine learning algorithms are highly complex and are difficult for human beings to understand. This raises an ethical issue as to whether the results of the INSPECTr tools will be properly comprehended by end-users. Misinterpreting what the tools are saying could have serious consequences for LEA investigations progress.

For NLP tools, it was determined that the confidence in models should be displayed with results. Further, partners agreed that the tools would need to be trained with models to reflect the situation the tools are expected to be used in (e.g. language in the context of criminal investigations), and that these would be documented to ensure traceability; this will continue after the project when new

models are used with the INSPECTr tools (as noted below, partners are exploring technology to ensure traceability; these are examples of the implementation of the Human agency, liberty and dignity, Transparency, Accountability, Non-instrumentalism, and Self-determination principles from section 3.2).

For the Cross-Correlation tool, it was recommended that the inputs should be masked so that users can understand impacts. However, it was noted that due to the complexity of the tools, and the different ways in which results will be displayed that a catch-all solution might be too difficult and, in any case, that this could only be fully considered toward the end of the project when the tools are nearer completion. Further work is therefore necessary on how the tool's results will be visualised and displayed, and how they this can be done in compliance with the requirements in section 3.2, particularly Human agency, liberty and dignity, and Non-instrumentalism.

Accountability

With respect to accountability, the need for adequate testing of algorithms was discussed. This is an ethical imperative as the impacts that could be created by different tools cannot be fully assessed or comprehended unless and until the tool capabilities are properly understood. It was determined that the already planned testing would be the best opportunity to test tools for bias, transparency, and explainability issues, and that the Quality Plan sets out the Software Testing Framework (therefore contributing to fulfilling the Human agency, liberty and dignity, Transparency, Accountability/Responsibility requirements in section 3.2).

4.2.1.3 Gender and AI – June 2021

In the workshop, the project considered gender in multiple ways, which are likely to interact[29]:

- Gender as bodily attributes (sex)
- Gender identity – a person's "felt, desired or intended identity"
- Perceived gender: the way in which a person is gendered and perceived by others and
- Gender roles: the (often gendered) behaviours and roles a person occupies.

Gender is an important issue for INSPECTr to address is that any of the above can be ways in which social power operates. A socio-technical system can have differential impacts upon people based upon any of those gender dimensions. Also, any system would be deployed into a world where gender is an important part of the social context of use.

Impacts can include:

- Lower quality of service for women and gender-nonconforming individuals
- Unfair allocation of resources, information and opportunities
- Reinforcement of existing harmful stereotypes and prejudices related to gender
- Derogatory and offensive treatment or reassure of already marginalised gender identities
- Detriments to physical safety and health hazards.[30]

[29] Keyes, O., May, C., & Carrell, A., "You Keep Using That Word: Ways of Thinking about Gender in Computing Research", Proceedings of ACM Human -Computer Interaction, Vol.5, No. CSCW1 Article 39, April 2021, https://ironholds.org/resources/papers/gender_multiplicity.pdf

[30] Smith, G. & Ishita, R., "When good algorithms go sexist: Why and how to advance AI gender equity", Stanford Social Innovation Review, 31 May 2021, https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity

Assumptions about gender as a fixed category

Issues with IT systems can occur when designers make unwarranted assumptions about what is "normal", either for a user, or for some aspect of the world a system is trying to capture. In the field of gender, c**ommon assumptions include**:

- Gender is binary, and all people can be categorised unproblematically as male or female
    - In reality, some people are intersex, some people are gender non-conforming, some people are transgender.
    - Several countries now allow people to legally register non-binary gender identities[31]
- Gender is consistent over a person's lifetime
    - (it isn't – 22 EU Member states and the UK have established clear legislation to allow individuals to go through legal gender recognition).[32]

The project team and external experts discussed the potential ethical harms that could be raised by conceptualising gender as a fixed category; for example, a person who is transgender or intersex might be misgendered and treated differently by a tools that cannot comprehend this reality.

With respect to the NLP tools, it was discussed that gender would be considered along with age, type of crime, and other relevant data to engage in victimology analysis for classification (and, potentially, some forecasting activities). However, this data would come from victims of crime who should be able to self-identify and this should be captured by the INSPECTr tools, so mis-gendering should not be an issue for NLP tools. Further, classification will come from statistical differences, which should be independent from the meaning of the text itself (these approaches contribute to fulfilling the Diversity, non-discrimination and fairness requirements from section 3.2).

An issue for the evidence exchange/query tools is that ways of representing gender are different for different LEAs, so there would need to be some form of standardisation for evidence requests when the tools is operational. Partners were recommended to work on this to avoid potential harms of wrongly categorising people. It was also noted that properly capturing gender information has operational benefits as inaccuracies in crime data analysis could mislead an investigation (these approaches contribute to fulfilling the Diversity, non-discrimination and fairness, Human agency, liberty and dignity, and Non-instrumentalism requirements in section 3.2).

Missing data with a gender skew

Gender data gaps have commonly been identified in areas of crime and violence against women. Crimes commonly experienced by women – e.g. sexual harassment in public places - tend to be under-reported in comparison with other crimes, are often poorly classified, or not recorded in official crime statistics.[33]

Where men are assumed to be representative of all people, and so only data on men is used to represent all people, this can lead to significant harms for women.[34] For the crime prediction tools, it was recognised that this would be difficult to deal with as the original data is inevitably inaccurate due to under-reporting; for example, LEAs do not have an accurate picture of domestic violence because many victims do not report such crimes, and so predictive tools are unable to provide an accurate forecast of domestic violence trends. It was agreed that due to the data quality issues of inaccurate or

[31] https://en.wikipedia.org/wiki/Legal_recognition_of_non-binary_gender

[32] European Commission Directorate-General for Justice and Consumers, *Legal gender recognition in the EU: The journeys of trans people towards full equality,* June 2020, https://ec.europa.eu/info/sites/info/files/legal_gender_recognition_in_the_eu_the_journeys_of_trans_people_towards_full_equality_sept_en.pdf

[33] Gardner, Natalie, Cui, Janqiang and Coiacetto, Eddor, "Harassment on public transport and its impacts on women's travel behaviour", Australian Planner, 54:1, 8-15, 2017.

[34] See, for example, Criado Perez, Caroline, *Invisible Women*, Penguin, London, 2019.

incomplete data, this tool is ethically problematic. As such, it is important to communicate the potential for inaccurate results to end-users so they can better understand the outputs they generate, and the project should not introduce any additional biases to make the situation any worse (an example of the Fairness and Accessibility principles in section 3.2).

Gender biases in machine learning

The world has gender discrimination, and this will make patterns in training data sources. Machine learning on these data sources can pick up these patterns.[35] There are now quite a large number of studies of gender bias within Natural language processing.[36] For example, research on natural language processing finds significant gender bias in how models view occupations.[37] Examples can include a computer vision application labelling a person in an image as a male because there is a computer in the background.[38] Problems from this include perpetuating damaging stereotypes in downstream applications.

As noted above, data quality issues of inaccurate or incomplete data with respect to gender can have a damaging effect on investigation progress, and on people who are directly affected by it. Here, however, the harm comes not from someone being specifically mis-gendered, but being wrongly associated with something due to gender stereotypes. For example, a machine learning algorithm that assumed all women were nurses and all men were doctors would be discriminatory and therefore unethical.

It was agreed that although some of the NLP tools were shown to have small distances for gendered topics, there was still a risk of residual bias from assumptions that are included in the training corpora and this could reinforce stereotypes. Therefore, technical partners were recommended to conduct a bias audit to determine the level of bias in the tools and how this could be mitigated (an example of the Human agency, liberty and dignity, Diversity, non-discrimination and fairness, Non-Instrumentalism, and Self-Determination requirements from section 3.2). Since this workshop, GN have evaluated the impact of biases in generic corpora and have found that gendered biases were not so significant that results would be impacted (see D4.4).

Further, it was discussed that the inclusion of a sentiment analysis tool was not recommended from an ethical perspective. The technology behind these tools is not proven and so should not be considered appropriate for use in a high-risk environment like an LEA investigation. For example, sentiment analysis tools have been found to rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases[39] As noted above, technical partner no longer intend to include this tool in the final platform for LEAs (an example of the Ought before can requirement from section 3.2). [40]

---

[35] Smith, G. & Ishita, R., 2021.

[36] Costa-jussà, M.R. An analysis of gender bias studies in natural language processing. *Nat Mach Intell* **1,** 495–496 (2019). https://doi.org/10.1038/s42256-019-0105-5, see also Leavy, S. Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. In: 1st International Workshop on Gender Equality in Software Engineering (GE). New York NY, US: ACM, 2018. pp.14-16.

[37] Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A., "GenderBias in Neural Language Processing", Preprint, https://arxiv.org/pdf/1807.11714.pdf

[38] See Burns, K., Hendricks, L.A., Darrell, T., Rohrbach, A., & Saenko. K., "Women Also Snowboard: Overcoming Bias in Captioning Models",.European Conference on Computer Vision(EECV'18). 2018; Simonite, T., "Machines taught by photos learn a sexist view of women", Wired, 21 August 2017, https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/.

[39] Park, J.,Shin, J., & Fung, P. "Reducing Gender Bias in Abusive Language Detection. In Empirical Methods of Natural Language Processing" (EMNLP'18) 2018; Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K., & Wang, W.Y., "Mitigating Gender Bias in Natural Language Processing: Literature Review, ACL 2009, https://arxiv.org/ftp/arxiv/papers/1906/1906.08976.pdf

[40] See D4.6, Section 2.2 'Ethics', p.8

Gender attribution/recognition

The discrimination risks of automatic gender recognition tools were discussed in terms of people being mis-gendered, particularly in unrecognised ways. Issues can occur with the use of automated gender attribution technologies – e.g., a computer vision system that automatically labels images as featuring a male or female. This causes issues when it is based upon assumptions about gender as essential, immutable, and fundamentally physiological (determined by physiological characteristics, ignoring the many social elements of gender, e.g., our assumptions about appropriate dress or hairstyle). As with the first issue, these technologies often fail for people with non-conforming gender identities. Misgendering can be psychologically harmful. has been criticised for consistently operationalising gender in a trans-exclusive way, and carrying disproportionate risk for trans people subject to it. Automated gender recognition is particularly likely to misclassify trans people therefore the deployment of such systems can create risks for trans people. [41] Non-binary people cannot be classified correctly in binary systems. Other computer vision tools might use gender assumptions as a 'short-cut' for quicker recognition. Where this is used, there is a potential for discriminatory harm.

It was discussed how gender is used as a way of narrowing down searches in the facial recognition tool. This uses gender to determine what a face looks like and prioritise that for an end-user. Whilst it is expected that an investigator would make a determination on gender, the tool can assist this. It was recommended that this tool is tested to determine what level of bias is included in this tool. The ability to search images by gender of people depicted in them is considered a necessary functionality by LEA partners. Therefore, work in INSPECTr should seek to mitigate secondary impacts from this (an example of the Human agency, liberty and dignity, Diversity, non-discrimination and fairness, and Self-Determination requirements from section 3.2).

4.2.2 Discussions following the INSPECTr PGA

During April 2021 (M20), the project held a general assembly (PGA) where all of the technical work so far was explained and the different tools that are being researched for inclusion in the INSPECTr platform were demonstrated. This event was a useful point to update the overview of the project as a whole, get into detail on specifics, and identify places where privacy and ethics support work was required. A number of issues were raised by Trilateral as the Ethics Manager in a WP8 meeting following the PGA and discussed with technical partners. The issues and how they are dealt with are described below:

Multiple gadgets in the INSPECTr platform displaying results simultaneously.

The version of the INSPECTr platform demonstrated at the PGA allowed a user to run several similar analysers/gadgets simultaneously, and display the results together. So, for example, it could display the results from several image recognition tools at the same time. It was recognised that this was useful as it could give confidence to end-users, but also posed a risk of automation bias and therefore a risk to human dignity. Giving numerical probabilities or confidence percentages could give end-users a better understanding of how the results were reached, what they mean, and what they should think about them. Whilst displaying a confidence Is preferable to showing an end result, partners should still consider how they could present the 'workings' behind the confidence to give end-users better understanding and context.

[41] Keyes, O., "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition, Proceedings of the ACM on Human-Computer Interaction, Voll.2, Issues CSCW, November 2018, pp.1-22, https://dl.acm.org/doi/10.1145/3274357

It was noted that gadgets which are classifiers would likely be giving results as a priority list e.g., age range detector showing images youngest to oldest. Image tools are primarily about filtering and prioritisation at the moment, although recognition tools could be included later. Prioritisation would be based on what the tool is most confident about, and, as noted above, results should be presented with a way of understanding them (both points above are examples of implementing inform strategy from section 2.3, and the Human agency, liberty and dignity and Non-Instrumentalism requirements from section 3.2).

With respect to thresholds for including images, they should be set lower than would normally be expected as we do not want LEAs to miss a CSAM offence, for example, meaning that it would be best to include some false positives so that end-users can deal with borderline issues. It was agreed that the tools should avoid sharp cut-offs as the use of categorisation in image tools is to help end-users find offenders/victims, not to determine who is an offender/victim, or who they are (an example of implementing the Individual, societal and environmental wellbeing and Net Benefit requirements from section 3.2).

Further, it was discussed that prioritisation can mean that people do not look further (e.g. few people go to the 2nd page of a Google search). However, bearing in mind the LEA context, the prioritised images are often enough for an investigation to establish an offence and officers don't need to delve too deep. It was agreed that the results of the Living Lab experiments would be useful in determining the thresholds that will provide the most value for end-users, and how the training curriculum will need to be adapted to take this into account (an example of implementing the Purpose requirement from section 3.2).

Much of the work toward improving prioritisation is 'behind-the-scenes' and so the importance of it could be lost if end-users are not aware of this and, for example, treat top prioritised result as definitive. Therefore, it is essential that the limitations of the tools are communicated and understood with respect to what the tool can do, what it is intended to do, and what it cannot do (an example of implementing the Self-Determination requirement from section 3.2).

Rationale for the inclusion of particular tools

It was noted that some of the tools that were demonstrated did not show high-levels of accuracy that might be expected or required in LEA investigations, and a potential ethical risk that end-use could be affected by poor quality results.

It was agreed that the Living Lab experiments will test whether the different tools are worth including, and decisions on inclusion of tools likely to be made in final reports. It is also important to note that some tools have already been abandoned, e.g., parts of the Pub-Sub no longer reflect the proposal as LEAs did not want the intended configuration (an example of implementing the Technical robustness and safety requirement from section 3.2).

Use-cases seeming 'solvable'.

It was discussed how the use-cases as currently described, had no dead ends or unexpected avenues for investigation, and if all aspects of the use-cases can be handled then the tools might not struggle, and this might give an optimistic view of the technologies. Which could potentially lead to harms later if end-users follow the results of an inaccurate tool.

Technical partners explained that the use-cases are already a technical challenge, especially the fraud case due to the volume of data and processing. However, if during the Living Labs the use-cases do seem too easy, they could be mixed up with different LEAs dealing with other use-cases to give a wider range of testing capabilities.

Partners agreed that no tool will work in all circumstances, and so it is important to make sure that tools are tested enough so that we know when it is broken, wrong, and failing gracefully, and that this would come from the results of testing. Bearing in mind the LEA context, many existing tools struggle in many situations, LEAs often use several tools simultaneously and the project would be adding to this to provide additional options (an example of implementing the Technical robustness and safety, and Purpose requirements from section 3.2).

Emotion/mood/sentiment analysis in images.

Following a demonstration of the emotion, mood, and sentiment analysis in the computer vision tools, it was discussed that there is research showing that emotion analysis can often struggle with recognising emotions.[42] This is especially true across cultural or ethnic groups. This, therefore poses a risk of investigators getting a wrong impression of suspects, witnesses, or innocent bystanders, and could wrongly divert an investigation toward an erroneous path.

Technical partners explained that the intention was for these tools to help with prioritisation; for example, an end-user could search for images of angry people which would be prioritised for end-users to go through manually. It was agreed that prioritisation could reduce potential harms in comparison to recognition, but the efficacy of this tool should be monitored as it would likely produce a flawed prioritisation. As noted above, this tool will no longer be included in the final platform for LEAs, due to these issues (an example of implementing the Human agency, liberty and dignity, Technical robustness and safety, Ought before can, and Purpose requirements from section 3.2).[43]

NLP work on linking online posts by the same author

It was recognised that NLP work in the project was supposed to be taking place on anonymous data. But, the linking of posts by a supposedly anonymous author presents a risk of de-anonymisation if enough information can be linked together.

The technical partner working on this tool (GN) explained that linking posts and identifying authors was part of their work. The data they are using is anonymised by persons outside of the project. If datasets are found not to be anonymous, then they will stop processing them (an example of the Proactive not Reactive, Privacy as the Default Setting, and Privacy Embedded into Design from section 2.1, the Minimise, Hide, Abstract, and Enforce strategies from section 2.3). The datasets are used to create models that will be useful for LEAs, if they determine that the models could be seen as including personal data then they will not be shared. This approach was considered to be acceptable by the Ethics Manger. However, risks associated with this tool beyond the project need to be considered in detail; for example, end-users discovering a disproportionate and unnecessarily large amount of private information on suspects, or misuse by non-democratic regimes (an example of implementing the Privacy and data governance, Transparency, Ought before can, and Purpose requirements from section 3.2).

Rationale for choosing blockchain

The need for secure storage of logs from the INSPECTr tools and platform was agreed by all. It was noted that the intended use of blockchain for information storage had positives in terms of security, but also accountability, traceability, and oversight. It was highlighted that blockchain is ideal for situation where there is little, or no, trust between parties to a transaction. With respect to the judicial

[42] Feldman Barett, Lisa, et al. "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements", Psychological Science in the Public Interest, Vol.20, Issue 1, 2019, p.
[43] See D4.6, Section 2.2 'Ethics', p.8

process, there might be little trust between prosecution and defence, or with cross-border evidence exchange, and so using blockchain can be useful to demonstrate that evidence has not been tampered with.

However, we can expect some level of trust between, and within, LEAs. If one, or a few, organisations have a lot of control over the blockchain then there is a risk of a "51% attack" where the previous blocks could be modified through collusion by multiple actors.[44] Further, starting a case again could create a new record of an investigation. If either of these two concerns were to be realised, then this would defeat the purpose of using blockchain as secure storage.

Following this discussion, whilst there is a private blockchain for storage within the platform (making sure that sensitive data here is not exposed outside LEA), there is expected to be secure storage of hashes on a public Ethereum blockchain to guard against manipulation of a private blockchain (an example of implementing the Privacy Embedded into Design, and Full Functionality principles from section 2.1, the Control, Enforce, and Demonstrate strategies from section 2.3, and Technical robustness and safety, Privacy and data governance, Transparency, Accountability/Responsibility, and Purpose requirements from section 3.2).

Potential use of location data

Until the PGA, there had not been substantial discussion of location data as a data source for the INSPECTr tools. As there are many potential ethical and legal concerns regarding location data, and what it can reveal about data-subjects. It was therefore questioned what the potential use of location data is expected to be in the INSPECTr tools and platform.

Technical partners explained that the use of location data would follow the use-case during the project. As the use-cases in the project are fictional, this should not raise many privacy concerns; even where real closed case data is used, location data has already been analysed as part of the original investigation and so should not contribute to further privacy harms (an example of implementing the Privacy as the Default Setting principle from section 2.1, the Minimise strategy from Section 2.3, the Privacy and data governance, and Ought before can requirements from section 3.2).

However, beyond the project, the way in which location data is used will be up to LEAs. As location data can be (special category) personal data, it is important to remember that the LED requires that personal data is only processed where there is a specific purpose, so this needs to be considered by LEAs and location data analytics should not be included unless required. There is an expectation that location data would be used to generate some sort of map on dashboards. However, it must be considered that there is a clear difference between engaging in a historical analysis of where people have been compared with a predictive analysis of where people are likely to be in future, and there are different ethical and legal issues for this. It was recommended that ethical and legal concerns regarding location data should be included in the project training curriculum (an example of the Technical robustness and safety, Privacy and data governance, and Ought before can requirements from section 3.2).

Linking of artefacts by the platform

During the platform demonstration, it was noted that when new data is added to the platform it is possible to see where information about individuals has previously been uploaded. This potentially creates a privacy risk where end-users might be able to see more private information about people involved in a case than they would need, or be able to see their appearance in other, unconnected,

[44] Goyal, Swati, "51% Attack Explained: The Attack on A Blockchain", FX Empire, 2021. Available at: https://www.fxempire.com/education/article/51-attack-explained-the-attack-on-a-blockchain-513887

cases. This might not create privacy harms where an LEA officer is already aware of links between suspects and other cases from their experience. But, if any officer can see an unnecessarily large amount of information about a suspect, witness, or victim, there is the potential for this to be unreasonably intrusive and create a privacy harm.

Technical partners explained that, during the current phase of research and development, the platform can link all data that has been uploaded to it. But, case management and segregation of data in use of the platform is something that is being discussed (an example of the Separation strategy in section 2.3). In the final platform, the data will be segregated so there will be less ability to link artefacts by default. There might be a capability to query other investigators on own node via Pub-Sub (i.e., looking for links within a single LEA). It was recommended that the linking of artefacts is limited in some way so that end-users do not receive all information without good reason (an example of the Ought before can requirement from section 3.2). It is also important to note that the LED requires classification and categorisation of the categories of people associated to an investigation, and this needs to be taken into account by the partners designing tools.

Training

Due to the complexity of the platform, and the individual tools, it was expressed that the importance of the training curriculum was worth considering continually. This is important from an ethics perspective as it should contribute to ensuring that end-users are provided with as full and as accurate training provision as possible. This is especially relevant in LEA situations as end-users might need to be able to explain how the tools were trained and how the tools work to a court room; an inability to do this would weaken a case and potentially cause additional harm to victims of crime. It was agreed that the training curriculum should be thorough and detailed and ensure that end-users are provided and explanation of how the tools will work. The Training needs analysis is being conducted in WP6, and will require input on ethics and privacy issues (an example of the Privacy Embedded into Design principle from section 2.1, the Control strategy from section 2.3, and the Privacy and data governance, and Accountability/Responsibility requirements from section 3.2.

4.2.3 Ethics reviews of deliverables

As noted in D8.3 (Second Report on Ethical Governance), the TRI Ethics Manager conducts ethics review of deliverables in addition to existing peer-reviews to ensure that ethical concerns raised during the tasks that are being reported are presented appropriately. Generally, this results in minor modifications to the text to ensure that ethical issues are adequately explained.

In some cases, however, these reviews can be opportunities for implementing Privacy and Ethics-by-Design approaches. An example of this is when reviewing D5.1, the TRI Ethics Manager noted that computer vision components were not intended to present a confidence with respect to why a tool made a particular recommendation. This could present an ethical issue if it means that end-users are not provided with enough information to get a full understanding of the results, as this could result in poorer decision-making for an investigation and the potential that offenders are free for longer than necessary or innocent people are wrongly suspected of criminality. Therefore, it was suggested that confidence figures should be added to the results of computer vision tools during an ethics review. Technical partners accepted this and modified the way in which the tool presents information to include a confidence figure to facilitate better understanding by end-users (an example of the Transparency requirement from section 3.2).

4.3 Principles into practice

The following table tracks how the principles discussed in the preceding sections are being implemented in INSPECTr.

Table 3: Tracking implementation of Privacy-by-Design and Ethics-by-design principles in INSPECTr

Principles	How does this manifest in INSPECTr?	Relevant Tasks or deliverables
Privacy-by-Design Foundational principles		
Proactive not reactive	Privacy was a consideration from the design of the project. There is a dedicated work package for ethical and privacy issues, and the project teams includes expertise on both social and legal aspects of privacy as well as engineers with privacy and information security expertise. The initial architecture plans reflected the importance of keeping different LEA data sources distinct from each other. Data protection and data governance considerations were made at the beginning of the project, and, where plans change, ahead of processing.	D2.2. Legislative Compliance relating to law-enforcement powers and evidence requirements WP8.
Privacy as the default setting	The query engine, which enables INSPECTr LEA users to make queries about evidence held by other LEAs requires explicit authorisation to perform a search. A requirement from D8.5 is that INSPECTr analysers (also known as "gadgets") which users can deploy to perform analytical operations on held data are disabled by default, and require initial enabling by an appropriate senior LEA official. The aim is that because these gadgets have the theoretical capacity to infringe upon privacy, their activation should be explicit. Further, tools should not be made available to LEAs who see no need for them. Within the project, use of non-personal, anonymised, and mock data is generally used, and (sensitive) personal data is only used where needed to reach the aims of the project.	D3.1 Security/Privacy preserving Publish-Subscribe Engine D3.4 Legislation Management tools for Data Exchanges
Privacy embedded into design	The query engine can be understood as privacy-preserving in that it allows for queries across different data sets without exposing those data sets	D3.1 Security/Privacy preserving

	to other parties. Whilst the origin of this requirement lies with the legal obligations upon law enforcement authorities, it can also be understood as working towards protecting the privacy of people included in those data sets. The systems for doing this involve translating organisational requirements into trust and privacy preserving business rules. Where INSPECTr tools are being built that query external data sources (e.g. online tools used for digital forensics) these are being designed so that the queries sent to these tools do not reveal the nature of the data being processed by INSPECTr. Where potentially sensitive data is being accessed, this is done in a privacy-respecting way by processing as little personal data as possible, or none if the aims of the task(s) can still be achieved. The training curriculum will include a dedicated module on ethical, legal, and societal issues, where privacy will be a key focus.	Publish-Subscribe Engine
Full functionality	INSPECTr is predicated upon facilitating the analysis of information that LEA's legally and appropriately hold. Before analysers (e.g. computer vision tools, any profiling tools) can be used, the data must first be ingested into the platform (where it will be logged), and stored securely. Further, secure storage of hashes for accountability purposes is also implemented. This is made more explicit in D8.2	D3.3 Data Ingestion Engine and API
End-to-End security	Security is of high concern to the potential LEA end-users of INSPECTr. Of paramount importance to the design / development goals of WP3, is that all tools will deliver automated compliance to the applicable legislation and will conform to governance and security requirements set in WP1 Using blockchain technology for integrity, validation and non-repudiation. Requirement #2 emerging from consultation with end-users and stakeholders is "Provide a secure **case management system** with administrative controls over access rights, sharing protocols, LEA	D1.2 Common Baseline Experimentation Environment and Detailed Requirements. D3.1 Security/Privacy preserving Publish-Subscribe Engine

	infrastructure integration, and the tools and related configurations available to authorised users." D8.5 Ethical, legal and societal requirements introduced requirements around using a 'traffic light protocol' to determine the security level of data that is accessed.	
Visibility and transparency	All investigative actions and operations using the INSPECTr platform (e.g. that a particular tool has been run on a particular set of evidence data, by a particular users) will be logged by default allowing retrospective accountability, at least within LEA themselves. Requirement #18 is "**Immutable recording** of the processing of the digital evidence that can be queried to attest of the **chain of custody** integrity." D8.5 introduced the requirement that "All AI systems (including systems labelling events and objects) must provide information on errors (e.g., false positives, false negatives) and other weaknesses (e.g., poorer performance on particular groups) in the model outputs to inform LEA decision making."	D1.2 Common Baseline Experimentation Environment and Detailed Requirements.
Respect for user privacy	This is a challenging principle for INSPECTr and reflects the consumer-facing origin of the 7 foundational principles of privacy-by-design. In many of the contexts envisaged in privacy-by INSPECTr's "users" will be LEA officers in analytic roles working in their professional capacity. As discussed above, there are strong reasons grounded in accountability and transparency to limit the privacy these users have in their use of the INSPECTr tool. The privacy focus is rather upon people who are in some way involved in the data being investigated or exchanged through the platform.	This report
ENISA privacy by design strategies		
Minimise	Guidance on the use of external sources and the queries to these sources to minimize collection of unnecessary personal data. Recommendation that web-crawling tools include filtering capabilities. Engaging in data minimisation strategies during the project.	T3.2

Hide	Filtering and prioritization of evidence data using machine vision tools, hides irrelevant data. Allowing queries across evidence databases held by different authorities without exposing the contents of those databases. Anonymising data where possible.	T4.4, T2.1
Separate	The node-based architecture of the INSPECTr platform maintains the appropriate separation between the personal data held by different LEA.	T3.1
Aggregate/Abstract	The exploratory work being done on crime forecasting tools involves aggregated data. Visualisation of results should consider how data can be aggregated or abstracted	T4.5, T4.2, T5.3
Inform	Not possible given the context of use. However, end-users should be informed of the privacy risks associated with their use of the INSPECTr tools.	See section 2.4
Control	Not possible given the context of use. However, the tools should be built in such a way as to provide sufficient control to end-users over data governance and security.	See section 2.4
Enforce	Information security requirements, Legal rules on data sharing and legislation management tools, recommendation that tools requiring activation by appropriate authority. Organisational requirements for data protection.	T2.1, T3.1, T3.4
Demonstrate	Full logging of all operations in the INSPECTr platform, and trust mechanisms using blockchain ledger allow for demonstration of compliance with policies and allows for detailed audit.	T3.4
SHERPA High-level ethical requirements		
Human agency, liberty and dignity	Facilitating understanding of how the tools work in the operational LEA context. Facilitating end-users to be able to exercise their own agency using accurate tools.	T1.3, T6.1
Technical robustness and safety	Ensuring that the INSPECTr tools are fit for purpose.	T3.1
Privacy and data governance	Minimising personal data use, and ensuring compliant processing. Ensuring end-users are aware of privacy risks	T1.3, T6.1

Transparency	Ensuring that the way the INSPECTr tools work is adequately understandable to users, and information needed for this is appropriately accessible.	T1.3, T5.1, T6.1
Diversity, non-discrimination and fairness	Engaging in bias audits of tools before exploitation. Building the query engine with capability to deal with information on persons who are gender non-binary. Building the INSPECTr tools to not use binary categorization of gender.	T1.4, T3.1, T4.3, T4.4 T6.4
Individual, societal and environmental wellbeing	Implementing appropriate thresholds for the different tools in INSPECTr to increase consideration of results by end-users.	T4.4
Accountability	Ensuring it is clear to end-users that they are responsible for how the tools work, and demonstrating this.	T1.3, T6.1
The Ethics Centre Principles		
Ought before can	Designing the INSPECTr project within the boundaries of research, and clearly avoiding direct law enforcement activities. Determining which tools and data usage would provide proper and appropriate utility to LEAs in use, and removing those that do not.	T1.3, T6.1
Non-Instrumentalism	Building the tools in a human-centric way.	WP2, 3, 4, 5
Self-Determination	Making ends-users aware of the (in)capabilities, and intended uses of tools. The use of different INSPECTr tools will be chosen by the LEAs, and the meaning of the outputs will be chosen by the end-users.	T1.3, T6.1
Responsibility	Ensuring it is clear to end-users that they are responsible for how the tools work, and demonstrating this.	T1.3, T6.1
Net Benefit	Building the INSPECTr tools to better facilitate the protection of victims of crime.	T4.4
Fairness	Building the INSPECTr tools to avoid adding additional biases.	WP2, 3, 4, 5
Accessibility	Not possible given the context of use. However, INSPECTr tools should be built to take account of vulnerable people whose data might be processed by the platform.	WP2, 3, 4, 5

Purpose	Ensuring that the INSPECTr project outputs contribute to fighting crime and terrorism.	T1.3, T1.4, T6.4

5 Conclusion

Across Sections 2 and 3, this deliverable has presented the concepts of Privacy-by-Design and Ethics-by-Design, including their historical background, and the current state-of-the-art. It has explained how, be combining these two design approaches, and doing so in the context of LEA technology development, the use of this process in INSPECTr is going beyond state-of-the-art.

In Section 4, key issues privacy and ethical issues that have been discussed in the INSPECTr project so far have been explained. Where design solutions have been found, or recommended, these have been provided. Some of the design solutions are ongoing tasks and will be finalised as the project progresses.