Building ethical, realistic, and compliant data for AI in security

Dr Henri BOUMA
Senior Scientist
TNO

Ezgi EREN
Doctoral Researcher
KU Leuven Centre for IT & IP Law

> Achieving this vision required more than collecting data. It meant constructing an entire ecosystem where datasets could be created, shared, and reused safely, transparently, and in alignment with GDPR and the EU AI Act.
>
> Dr Henri Bouma, TNO

Responsible data: the cornerstone of trustworthy AI From the outset of STARLIGHT, the project's dataset development efforts focused on one of the initiative's most complex challenges: creating realistic and representative datasets for training, testing, and benchmarking Artificial Intelligence (AI) tools in a legally, ethically, and socially compliant manner.

Achieving this goal required not only the technical precision involved in the creation and curation of data but also the establishment of a secure, transparent, and responsible data ecosystem capable of supporting advanced AI development without compromising privacy, rights, or trust.

The consortium approached this mission by focusing on three interconnected objectives: ensuring realistic and representative data, protecting sensitive information through anonymisation or synthetic generation, and preventing bias in dataset creation.

These objectives formed the foundation for a comprehensive strategy combining data collection, generation, and rigorous balancing.

From concept to compliant practice Guided by continuous feedback from Law Enforcement Agencies (LEAs), the team created datasets that reflect realistic operational conditions while maintaining full compliance with ethical and legal standards. Privacy-preserving mechanisms were embedded throughout the process, including anonymisation, synthetic data generation, and secure handling protocols. These measures allowed AI tools to be developed and tested responsibly without exposing personal data.

In total, 16 tools were developed within this effort: five for data collection and annotation, three for anonymisation, and eight for synthetic data generation. Thirteen of these are available in the STARLIGHT repository for the whole consortium, while three can be made available to European LEAs upon request.

This work also included the assessment of more than 200 datasets for legal and ethical aspects. Ninety datasets were uploaded to the repository, including 58 existing or public datasets and 32 new ones, covering collected, anonymised, and generated data.

No dataset was used within STARLIGHT without first undergoing a thorough legal and ethical assessment procedure. Each dataset was documented, reviewed for Ethical, Legal and Social Aspects (ELSA) by partner KU Leuven, and approved following data protection best practices.

This process ensured that every dataset entering or leaving the project complied fully with the General Data Protection Regulation (GDPR) and STARLIGHT's internal ethics framework.

Lessons for future research The project highlighted a key lesson for future R&D: arranging safeguards and approval for a large initiative aiming to create many datasets and AI applications for multiple use cases becomes disproportionately complex compared to the effort for research and development. For this reason, future AI research and development projects focusing on a single application with one clear purpose may ensure that ELSA governance remains more manageable.

This hands-on experience also proved invaluable in resolving crucial mismatches between engineering and legal understandings of personal data, particularly concerning public datasets and the limits of data processing.

Ultimately, this effort delivered a robust data ecosystem that supports innovation, ensures compliance, and contributes directly to STARLIGHT's mission of enabling transparent, safe, and trustworthy AI for European law enforcement.

16

Data-related tools developed

For data collection, anonymisation, and synthetic data generation, delivering practical solutions to create, protect, and enhance data for AI development.

200+

Datasets evaluated

Reviewed and validated for ethical and legal compliance, ensuring every dataset meets GDPR and STARLIGHT's ethics framework before use.

90

Datasets published

Shared through the STARLIGHT repository for LEA research and testing, providing secure, accessible resources to support responsible AI innovation.