TAP
Technology
Advisory Panel
IPCO

Proportionality Guidelines:
Privacy intrusion in data collection and analytics

24 November 2022

Necessity and proportionality are integral to the justification of warrants. These guidelines deal specifically with proportionality and assume the case for necessity and resources have already been made.

They provide guidance on factors to take into account that may affect privacy intrusion. Decision makers have discretion when balancing factors and coming to a final decision; therefore, there is scope to depart from these guidelines where appropriate. Guidelines are guidelines not tramlines.

Part A gives three lists (non-exhaustive) of factors to consider in assessing proportionality; while resources are typically not a concern for IPCO, they are included for completeness as they may be referred to in a case. Part B introduces and gives examples of trade-offs.

Part A: Factors

Factors are grouped into three categories: data collection and analytics, privacy intrusiveness, and resources. Within each category, each factor is described by high-level descriptors, on the left, and more detailed descriptors in the right. Factors in bold and starred (*) would usually carry more weight.

1. Factors relevant to data collection and analytics

Value	
Timeliness and need *	• gravity and extent of (potential) crime or harm • public interest • urgency of need
Function *	• for analysis of the data on its own • to enrich existing data • to become enriched by existing data • for training sets for use in machine learning algorithms in established tools • for use in development or enhancement of a new capability or tool, which may be a prototype
Relevance and marginal benefits *	• to given investigation(s) • to other data available
Impact of time and place	• dependencies such as when and where data were collected

Type of data or collection method	• new or existing type of data • new, more accurate, or existing collection method

Volume	
Amount *	• fixed and known before collection • unknown but can be approximated • granularity and uncertainties of approximations including dependencies
Frequency	• one-time collection • repeated collection, how many times and at which intervals • continuous collection, for how long • how does the amount of data held vary over time

Data Management	
Storage	• where, how, and under whose authority • length of time planned retention, for which parts • security of access and resilience to corruption or loss
Deletion and manipulation	• plans and mechanisms for indexing, deletion and/or putting beyond use, redaction, and abstraction

Analysis	
Human and/or machine inspection *	• uncertainty (false positives/negatives) thresholds for human and machine inspection • risks of bias for human and machine inspection • human only inspection is possible of entire data set • machine only inspection is possible of entire data set • primary analysis by machine inspection to extract set for secondary analysis by human inspection

Alternatives	
What other methods have been considered	• if they have been implemented successfully, why are they not employed now • if they have not been implemented successfully, why not • opportunity cost - what will be lost by implementing this method over others • efficiency and effectiveness of proposed method vs. alternatives

2. Factors relevant to intrusiveness

Privacy Intrusion	
Type *	• degrees of foreseeable, targeted, collateral, and privileged intrusion – how many individuals • their interrelationships and dependencies
Sensitivity *	• degree of sensitivity of the data collected and/or what will be revealed through subsequent analytics

Scaling	• how the intrusion scales from individuals to different populations e.g. multiplicative, additive, constant • how the intrusion affects a community defined by a characteristic
Access	• breadth of people (e.g. analysts) and systems that will have access either directly to the data collected or indirectly via analytical tools • breadth of people (e.g. analysts, colleagues, managers) who will have access to reports that refer to the data

3. Factors relevant to resources

Resources required for
• data collection • computation of enrichment/inference with other data • ownership including maintenance, utilisation, security, and deletion • all of the above for any alternatives

Part B: Balancing factors

Balancing is a judgment. The primary consideration is privacy intrusion, but the resources required may also be relevant to the public authority, e.g. a new data collection method may offer only a minor improvement in intrusion but require significantly more resource.

Balancing may be informed by trade-offs and dependencies that exist:

- *within* collection and analytics factors
- *within* intrusiveness factors
- *between* collection and analytics factors and intrusiveness factors
- *between* resource factors and intrusiveness factors.

Trade-offs within resource factors are not usually a concern for IPCO.   Factors involved in trade-offs may range from high-level, e.g. *value*, to more detailed, e.g. *urgency*. It may be helpful to visualise trade-offs and consider the shape of the curve, points of inflection, and marginal benefit at particular points. Example trade-offs are given in the Annex.

Trade-offs within data collection and analytics

- value vs volume
- uncertainty thresholds in machine inspection vs volume requiring human inspection

Trade-offs within intrusiveness

- collateral vs targeted intrusion
- sensitivity of data vs access to data and reports

Trade-offs between data collection and analytics and intrusiveness

- urgency vs sensitivity
- relevance vs collateral intrusion

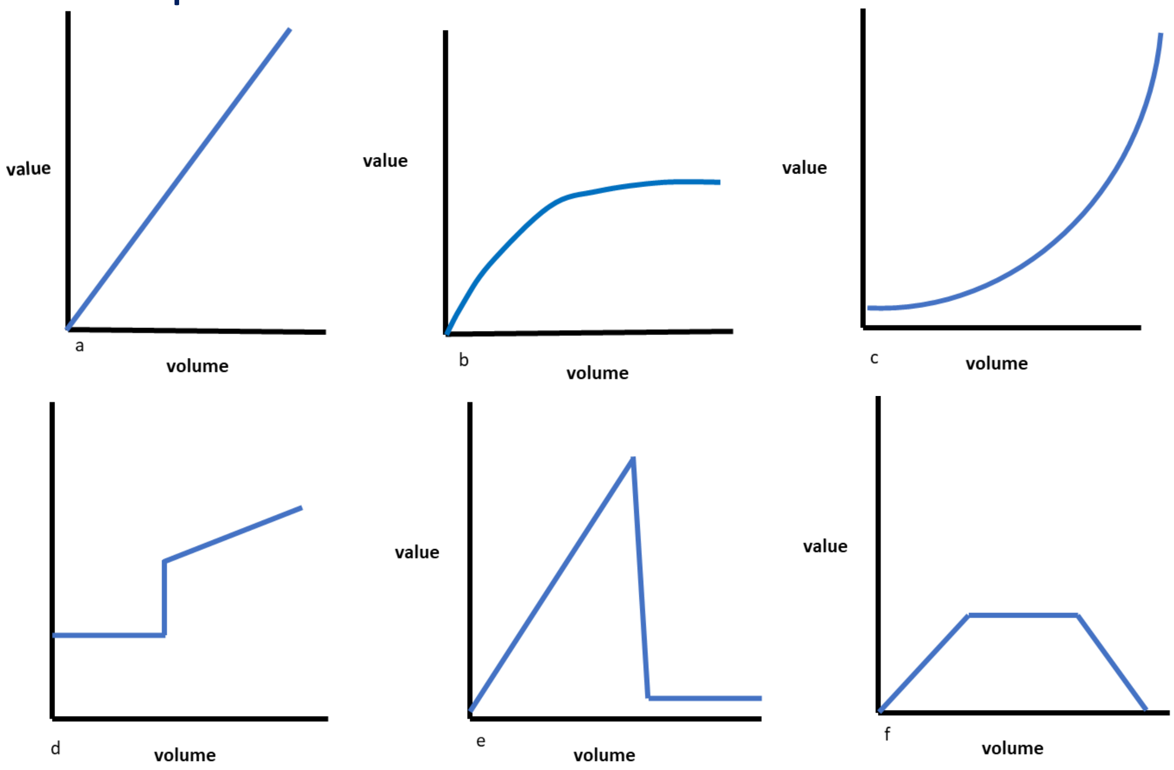
Trade-offs between resources and intrusiveness

- collection resources vs sensitivity
- ownership resources vs breadth of access

Annex: Visualising trade-offs

A (non-exhaustive) set of *contrived* exemplars illustrate a variety of possible curves and trade-offs.

1. Example curves for value vs volume



a) The relationship is *linear*.
Example: number plates seen at a traffic light on successive days.

b) Over time the value approaches an asymptote (a constant that it never exceeds), which means that *from some point onwards*, the *marginal benefit* of additional volume is very *small*.
Example: Collecting day and month of birth from people entering a building. There is an upper limit on the number of distinct values that can be collected and so after all those have been collected, no further new values can be collected, regardless of who enters the building.

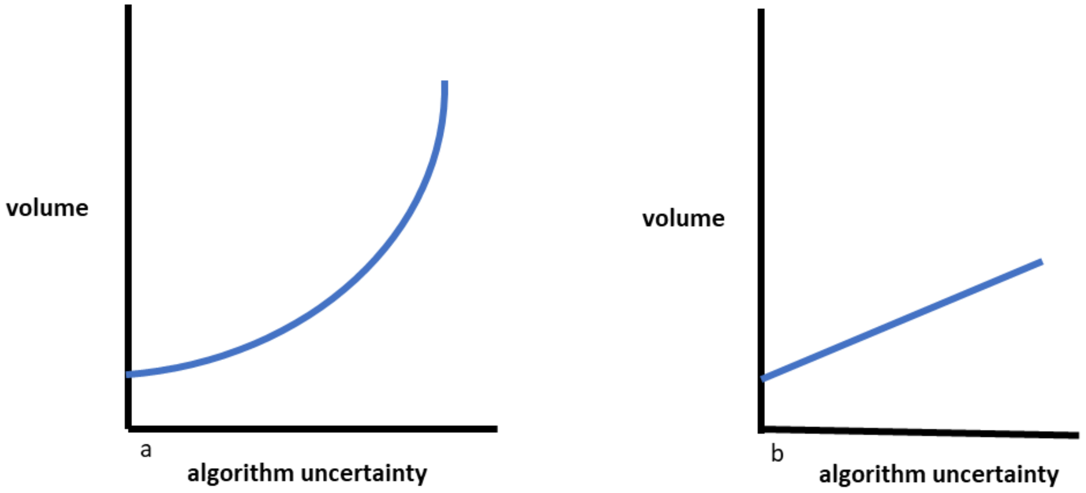
c) The value rises *exponentially*.
Example: Collecting information about contacts of a person of interest.

d) The value is *constant* for volumes up to a certain size, and then it jumps to higher value and continues to rise *linearly*. So, the there is *no marginal benefit* of additional volume until that amount has been reached.
Example: At least 100 data points are required before more valuable inferences can be made about the data.

e) The value rises *linearly* until a certain amount, after which it drops to a very low and *constant* value. So, the value has *decreased dramatically* after a certain amount.
Example: The collection method doesn't work properly in a crowd above a certain size.

f) The value rises linearly until a certain amount, after which it stays constant until there is so much volume that the value begins to decline and continues to do so.
Example: Collecting sets of phone numbers and examining their intersections. Starting from an initial set, as we collect further sets and examine their intersections, we gain value, but, without appropriate capability, we reach a point where each additional set does not bring in any new insight, and then because each set contains some element of randomness (i.e. numbers not of interest) each additional set actually decreases the insight gained.

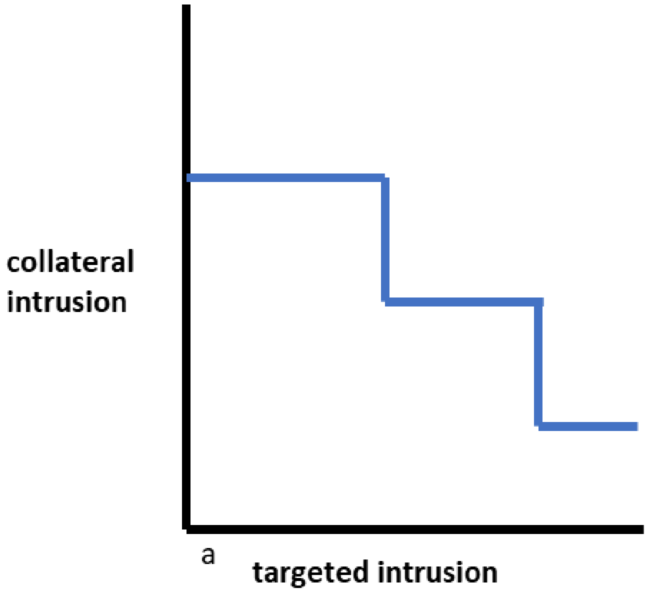
2. Example curves for algorithm uncertainty vs volume for human inspection (machine inspection vs human inspection)



a) The relationship is *exponential*.
Example: A ML algorithm for movement detection. As the uncertainty of the algorithm increases, the volume of results that require human checking increases exponentially. This means that very quickly human inspection of the results becomes intractable. Note that even when there is virtually no algorithm uncertainty, *some* results require human inspection.

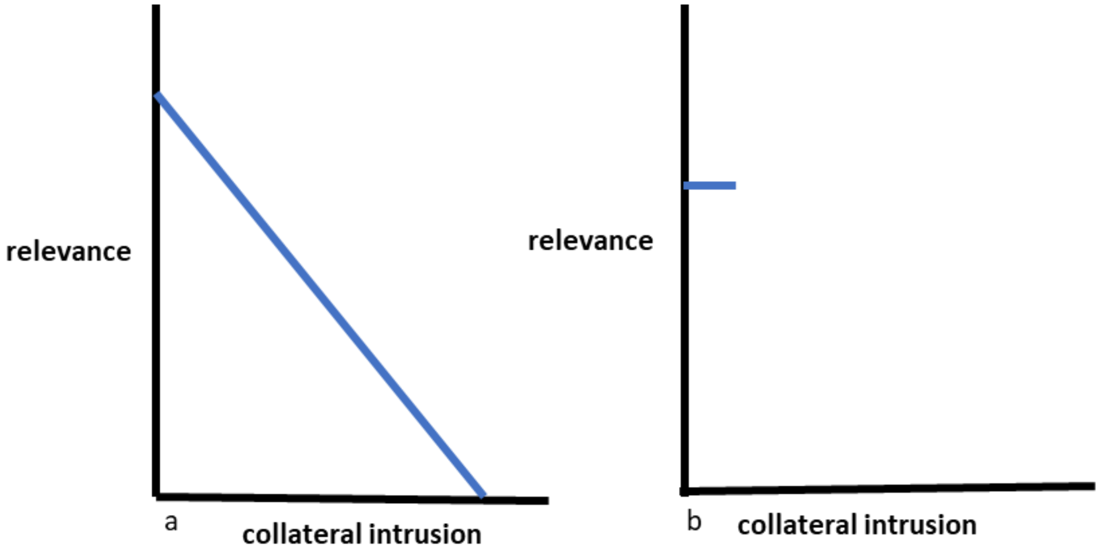
b) The relationship is *linear.*
Example: A different ML algorithm is less sensitive to uncertainty in the volume of results requiring human inspection, so the volume only increases linearly with the uncertainty. Note again even when there is virtually no algorithm uncertainty, *some* results require human inspection.

3. Example curve for collateral intrusion vs targeted intrusion



a) The relationship is *step-wise*.
Example: Analysis of phone calls made from a device. There may be collateral intrusion concerning anyone who uses that device. But over time, as more information is gained and analysed about the subject of interest, more calls to numbers that are not of interest become excluded, reducing collateral intrusion.

4. Example curves for relevance vs collateral intrusion



a) The relationship is a negative correlation.
Example: A camera on a building captures and stores images of everyone who comes to the door. If we are only interested in one subject and their associates, then as the collection increases in size its total relevance decreases while the collateral intrusion increases. In

other words, we may collect lots of information that is not very relevant and is collaterally intrusive.

b) The relationship is *constant*.
Example: The camera system immediately does not store images that do not match a set of subjects of interest. This means the data kept always has high relevance, and collateral intrusion hardly increases.