ASPEN TECH POLICY HUB

POLICY

ALONI COHEN

Rethinking Probabilistic Identifiers for the CCPA

The California Consumer Privacy Act (CCPA) – the nation's most expansive data privacy law – will go into effect in 2020. It governs the ways companies can use the personal information of consumers, and grants individuals certain rights to data that companies have collected about them. Before CCPA takes effect, California's Attorney General will issue regulations to clarify parts of the law, and updating the definition of personal information in particular.

One particular area where the legislative language is vague and flawed is with respect to probabilistic identifiers, which are a specific type of personal information. Understanding the meaning of probabilistic identifiers is critical to understanding the scope of CCPA. The Attorney General has the power and mandate to clarify the meaning of the term through regulation, providing an opportunity to make needed changes to ensure that the Act accomplishes its intended goals

In order to reduce uncertainty during transition to CCPA enforcement, the Attorney General's initial regulations on CCPA should provide that: 1) Probabilistic identifiers can take many forms; 2) The power of the data, not its form, is important; and 3) Probabilistic identifiers might not consist of personal information.

PROBABILISTIC IDENTIFIERS CAN TAKE MANY FORMS

Probabilistic identification is used to track and identify users online by making best-guess inferences using data that is not traditionally considered personally identifying. For example, a company can identify an individual user with a high degree of certainty by collecting as few as four or five non-personally identifiable pieces of information about a user – like their neighborhood, gender, ethnicity, age and which devices they use.

The data used to probabilistically track users online comes in many forms, including those described below. The Attorney General's regulations should clarify that all of these forms are covered by the law. Doing so would further clarify that the regulation covers datasets that have already been shown to enable probabilistic identification.

DATA SETS THAT ENABLE PROBABILISTIC IDENTIFICATION

A collection of information about multiple consumers or devices from which a probabilistic identifier may be inferred (e.g., raw data, device network).

A digital application or interface from which a probabilistic identifier may be inferred (e.g., interface to device network).

A collection of multiple attributes or other data about an individual consumer or device (e.g., inferred profile, information used for browser fingerprinting).

Any other information that can be used for probabilistic cross-device tracking or device-fingerprinting.

THE POWER OF THE DATA, NOT ITS FORM, IS IMPORTANT

The definition of "probabilistic identifier" is constructed inconsistently from the other definitions of information and identifiers in CCPA. The Attorney General should clarify that a "probabilistic identifier" is information which can be used to identify or recognize a consumer or device, rather than the identification of the consumer or device itself.

Probabilistic identifier is currently defined as the "identification of a consumer." This focuses the definition on the form of the data, not its power; what has been done with the data, not what can be done with it. Data may be treated differently than inferences drawn from it.

The authors of the CCPA recognized this distinction in the definitions of "personal information" ("information that identifies…"), "unique identifier" ("a persistent identifier which can be used to recognize…"), and "deidentified" ("information which cannot reasonably identify"). The difference is one of the many inconsistencies born of CCPA's frenzied passage and should not be considered a sign of legislative intent.

ASPEN TECH POLICY HUB

POLICY

ABOUT THE HUB

The Aspen Tech Policy Hub is a Bay Area policy incubator, training a new generation of tech policy entrepreneurs. We take tech experts, teach them the policy process, and support them in creating outside-the-box solutions to society's problems.

The Aspen Institute
2300 N St. NW, Suite 700
Washington, DC 20037
202 736 5800

THE ASPEN INSTITUTE

PROBABILISTIC IDENTIFIERS MIGHT NOT CONSIST OF PERSONAL INFORMATION

The Act suffers from a circular definition of "probabilistic identifier". It is a sub-type of personal information, but must be based on "categories of personal information." This creates an ambiguity that threatens to strip probabilistic identifiers of all meaning.

Taking the circular definition literally, the probabilistic identification of a consumer based on a collection of many discrete data points may not be considered a "probabilistic identifier" if each of the data points separately do not rise to the level of personal information. It is hard to believe that the legislature intended this interpretation.

The Attorney General should clarify that – while based on categories of information included in, or similar to, the categories enumerated in the definition of personal information – probabilistic identifiers may consist of a collection of multiple pieces of information that do not separately constitute personal information or unique identifiers.

PROPOSED REGULATORY LANGUAGE

Current language from the Act

> *1798.140(p) "Probabilistic identifier" means the identification of a consumer or a device to a degree of certainty of more probable than not based on any categories of personal information included in, or similar to, the categories enumerated in the definition of personal information.*

The only other occurrence of the term probabilistic identifier in the Act is in the definition of "unique identifier," which is one type of "personal information." The Attorney General has the power and mandate to update these definitions through regulation. Relevant excerpts from the legislation are included at the end of this document.

Proposed regulatory language

> *"Probabilistic Identifier" means information that can be used to identify a consumer or device to a degree of certainty more probable than not, and that is based on categories of information included in, or similar to, the categories enumerated in the definition of personal information.*
>
> *Probabilistic information may consist of a collection of multiple pieces of information that are not individually personal information, unique identifiers, or probabilistic identifiers. Probabilistic identifier includes, but is not limited to, the following [if it can be used to recognize a consumer or device to a degree of certainty more probable not and that is based on categories of information included in, or similar to, the categories enumerated in the definition of personal information]:*
>
> *(a) A collection of multiple attributes or other data about an individual consumer or device, including, but not limited to, information collected over a period of time;*
>
> *(b) A collection of information about multiple consumers or devices from which a probabilistic identifier may be inferred;*
>
> *(c) A digital application or interface from which a probabilistic identifier may be inferred, including, but not limited to, a cross-device or identity graph or an interactive statistical tool; and*
>
> *(d) Other information that can be used for probabilistic cross-device tracking or device- or browser-fingerprinting.*

DISCUSSION OF PROPOSED REGULATORY LANGUAGE

Current language

> *"Probabilistic identifier" means the identification of a consumer or a device to a degree of certainty of more probable than not . . .*

Proposed change

> *"Probabilistic Identifier" means information that can be used to identify a consumer or device to a degree of certainty more probable not, . . .*

The current definition of "probabilistic identifier" is constructed inconsistently from the other definitions of information and identifier in the legislation. Probabilistic identifier is defined as the "identification of a consumer."

This focuses the definition on the *form* of the data, not its *power*; what has been done with the data, not what can be done with it. This is very different from and inconsistent with the definitions of "personal information" ("information that identifies…"), "unique identifier" ("a persistent identifier which can be used to recognize…"), and "deidentified" ("information which cannot reasonably identify").

The regulation should clarify that a "probabilistic identifier" is a type of information which can be used to identify or recognize a consumer or device, rather than the identification of the consumer or device itself.

Current language:

> *. . . based on any categories of personal information included in, or similar to, the categories enumerated in the definition of personal information.*

Proposed change:

> *. . . and that is based on categories of information included in, or similar to, the categories enumerated in the definition of personal information.*

> *Probabilistic information may consist of a collection of multiple pieces of information that are not individually personal information, unique identifiers, or probabilistic identifiers. . . .*

The current definition is circular. Probabilistic identifiers are a sub-type of personal information (1798.140(o)(1)(A) and 1798.140(x)). As passed, probabilistic identifiers must be based on certain "categories of personal information." "Probabilistic identification" is a term typically used in industry to describe the identification of an individual based on many different types of information (e.g., operating system, browser, IP address, websites visited). Even though these data are in the "categories … enumerated in the definition of personal information," they arguably do not individually constitute personal information because they cannot by themselves be linked to a particular consumer or device.

This interpretation would render the notion of probabilistic identifiers toothless, and it is hard to believe that the legislature intended this reading. The suggested language changes the wording in the first sentence to clarify, and adds a new sentence to make this point explicitly.

Proposed addition:

> *Probabilistic identifier includes, but is not limited to, the following [if it can be used to recognize a consumer or device to a degree of certainty more probable not and that is based on categories of information included in, or similar to, the categories enumerated in the definition of personal information]:*
>
> *(a) A collection of multiple attributes or other data about an individual consumer or device, including, but not limited to, information collected over a period of time; . . .*

A large collection of data can very often be used to single out an individual. Examples include:

- Browser fingerprinting:[1] Details about computer hardware and software that can be used as a "fingerprint" online. This data is collected all at once.
- Browsing history: What websites you visit or apps you use
- Location or IP address over time: Where you are at various times of day.

1 http://panopticlick.eff.org

- Netflix Prize dataset:[2] Some Netflix users rate movies after they watch them. This has been used to identify them in publicly released datasets.

The first three are widely in the advertising industry and are the most common sources of data used for probabilistic cross-device identification.

Proposed addition:

> *(b) A collection of information about multiple consumers or devices from which a probabilistic identifier may be inferred; and . . .*

Data that underlies the probabilistic identification may be collected en masse and only later analyzed to extract individualized information. As the wellspring of probabilistic identifiers, this data is itself a probabilistic identifier. Examples include:

- The not-yet-identified raw data about the browsing activity thousands of users and devices that can be used to probabilistically identify individuals.
- The data releases published by the US Census Bureau based on the 2010 Decennial Census, which have been shown to allow the accurate reconstruction and reidentification of tens of millions of Americans.[3]

Proposed addition:

> *(c) A digital application or interface from which a probabilistic identifier may be inferred, including, but not limited to, a cross-device or identity graph or an interactive statistical tool.*

Within the advertising (and other) industries, access to data is often through some sort of application or interactive interface that gives some sort of access to underlying data. These interfaces allow one to make requests for information and automatically generates responses. Examples include:

- Cross-device graph API: A bedrock of the advertising industry is a continuously updated graph of the connections between various devices (e.g., this phone and this computer are owned by the

2 https://arxiv.org/abs/cs/0610105
3 https://twitter.com/john_abowd/status/1114942180278272000?lang=en

same person). Companies construct this graph by making inferences from various data sources. Other companies then pay for some type of access to this graph in order to improve or measure the effectiveness of their advertising campaigns. An interface might allow access to individualized information or only provide aggregated statistical information. However, even aggregated statistical information can often allow information about an individual to be inferred.

- Interactive government (or business) statistics: Governments sometimes make webpages which allow users to discover statistics based on confidential government surveys (e.g., Australian Bureau of Statistics Table Builder, Israeli Central Bureau of Statistics). Similar technology is in development in private industry (e.g., Diffix). These systems sometimes allow the underlying data to be reconstructed by querying many statistics about the underlying data and then probabilistically reconstructing it.[4]

This language is meant to clarify that these interfaces, which may reveal or communicate probabilistic identifiers, are themselves probabilistic identifiers.

Proposed addition:

(d) Other information that can be used for probabilistic cross-device tracking or device fingerprinting.

Probabilistic identification is most often used in the ad tech industry to refer to probabilistic cross-device tracking and device fingerprinting. Both of these techniques exist solely to infer the identity of an individual (or device) when identification cannot be done with certainty. Information used for these purposes are by their very nature probabilistic identifiers.

Without regulatory clarification, the meaning of probabilistic identifiers under CCPA will remain unclear. The proposed language closes potential loopholes that would allow probabilistic online tracking to escape CCPA regulation.

4 https://www.haaretz.com/surveys-not-as-anonymous-as-respondents-think-1.5288950, https://arxiv.org/abs/1902.06414 https://arxiv.org/abs/1810.05692

ADDENDUM:
ATTORNEY GENERAL'S POWER AND MANDATE TO UPDATE THE ACT'S DEFINITIONS THROUGH REGULATION

The only occurrence of the term probabilistic identifier in the Act (aside from the definition itself) is in the definition of "unique identifier," which is one type of "personal information." The relevant definitions are (emphasis added):

> *1798.140(o) "**Personal information**" means information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household. Personal information includes, but is not limited to, the following if it identifies, relates to, describes, is reasonably capable of being associated with, or could be reasonably linked, directly or indirectly, with a particular consumer or household:*
>
> *(A) Identifiers such as a real name, alias, postal address, **unique personal identifier**, online identifier, Internet Protocol address, email address, account name, social security number, driver's license number, passport number, or other similar identifiers. . . .*
>
> *1798.140(x) "**Unique identifier**" or "Unique personal identifier" means a persistent identifier that can be used to recognize a consumer, a family, or a device that is linked to a consumer or family, over time and across different services, including, but not limited to, a device identifier; an Internet Protocol address; cookies, beacons, pixel tags, mobile ad identifiers, or similar technology; customer number, unique pseudonym, or user alias; telephone numbers, or other forms of persistent **or probabilistic identifiers that can be used to identify a particular consumer or device**. . . .*

The Attorney General has the power and mandate to update these definitions through regulation. The Act provides that (emphasis added):

> *1798.185 (a) ...[T]he Attorney General shall solicit broad public participation and adopt regulations to further the purposes of this title, including, but not limited to, the following areas:*
>
> *(1)* ***Updating as needed additional categories of personal information*** *. . . in order to address changes in technology, data collection practices, obstacles to implementation, and privacy concerns.*
>
> *(2)* ***Updating as needed the definition of unique identifiers to address changes in technology****, data collection, obstacles to implementation, and privacy concerns, . . .*
>
> *(b) The Attorney General may adopt additional regulations as necessary to further the purposes of this title.*