



Comissão de Ciência, Tecnologia e Informática (CCTI)

## Audiência Pública

(Requerimentos nº 01, 14, 29, 52, 71, 79 e 134 de 2021)

Prof. Thiago Tavares  
Presidente da SaferNet Brasil

Brasília, 18 de novembro de 2020

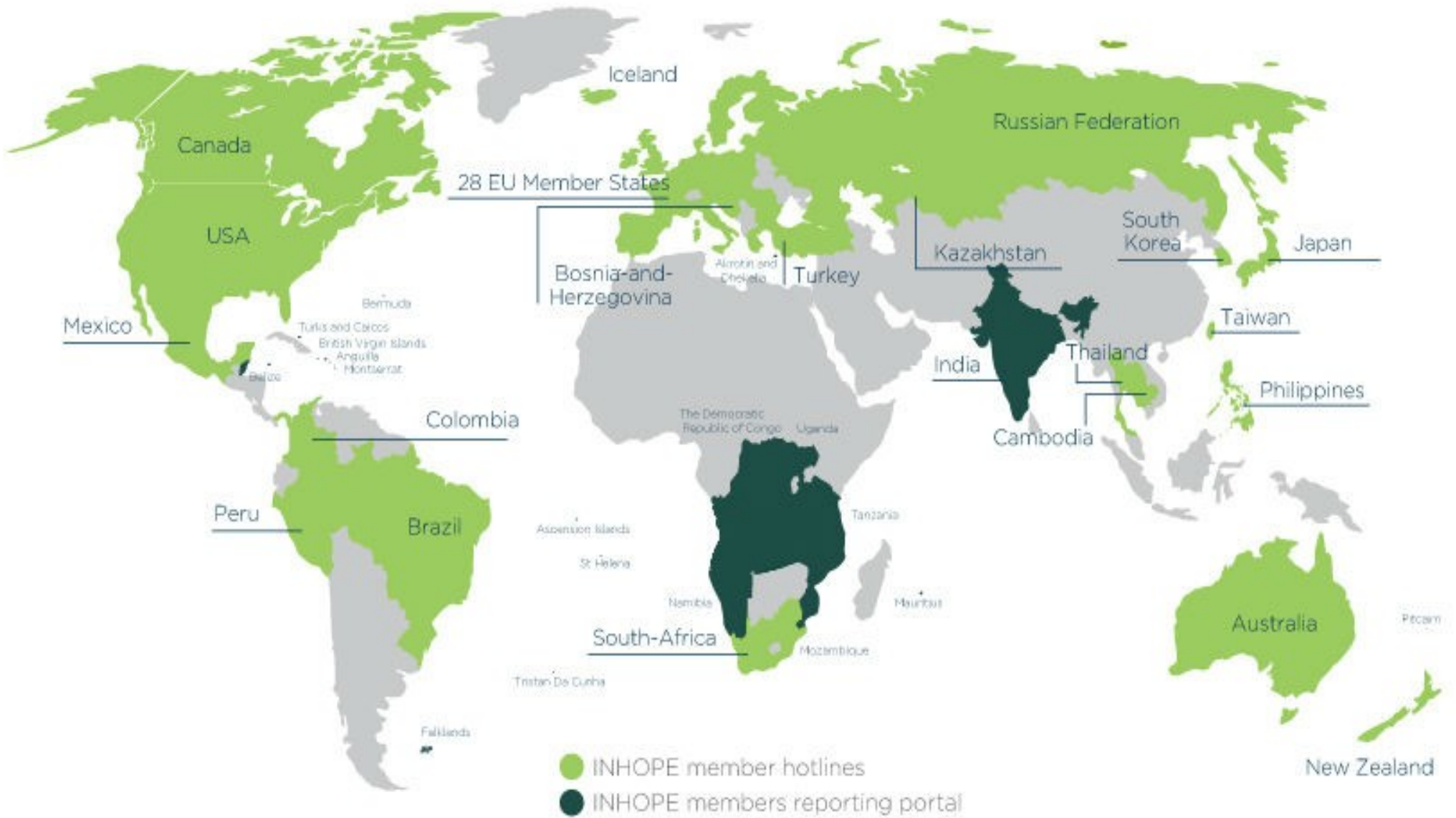
## Resumo da minha atuação profissional e acadêmica relevante para o debate de hoje:

- Representante da sociedade civil no Conselho Consultivo sobre Internet e Eleições do TSE, durante as gestões dos Exmo. Srs. Ministros Gilmar Mendes, Luiz Fux e Rosa Weber (2017/2028)
- Representante da SaferNet Brasil no programa de enfrentamento à desinformação do TSE durante a gestão do Exmo. Sr. Ministro Luis Roberto Barroso (2020/2021)
- Liason partner no Brasil do Partnership for Countering Influence Operations (PCIO), instituído pelo Carnegie Endowment for International Peace (EUA) com o objetivo de promover uma comunidade internacional multidisciplinar de especialistas que trabalham para entender as operações de influência em ambientes digitais, incluindo campanhas massivas de desinformação (2019 - ...)
- Representante (membro titular) eleito e reeleito pelo segmento do terceiro setor para representar a sociedade civil no Comitê Gestor da Internet (CGI.br), entre 2014 e 2020
- Representante da sociedade civil no Conselho de Administração do Núcleo de Informação e Coordenação do Ponto Br (NIC.br), braço executivo do CGI.br (maio/2017 a maio/2021)
- Presidente da INHOPE Foundation e Diretor da INHOPE Association (2014/2016), associação internacional com sede em Amsterdam/Holanda e membros em 43 países, instituída em 1999 como parte do Safer Internet Program da Comissão Europeia com o objetivo de apoiar a rede de canais de denúncia no combate a disseminação de material de abuso sexual infantil online
- Professor de Direito e Tecnologia desde 2005, e especialista convidado por diversas empresas, universidades, órgãos públicos e think tanks, no Brasil e no exterior, para opinar sobre temas relevantes envolvendo o aperfeiçoamento da legislação brasileira nas áreas de Segurança Digital, Direitos Humanos e Governança da Internet
- Fundador e Presidente da SaferNet Brasil (2005 - ...)





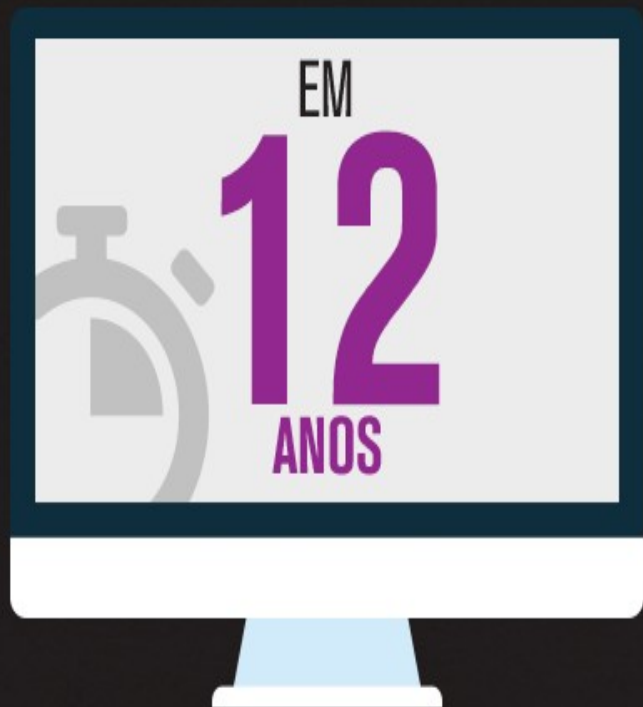
**Safer**  
**net**



3.925.405

DENÚNCIAS ANÔNIMAS

RECEBIDAS DO  
CANAL DE  
DENÚNCIA



701.224

PÁGINAS (URLS) DISTINTAS

9

IDIOMAS

94.155

HOSTS DIFERENTES

56.416

NÚMEROS IPS DISTINTOS

101

PAÍSES

5

CONTINENTES

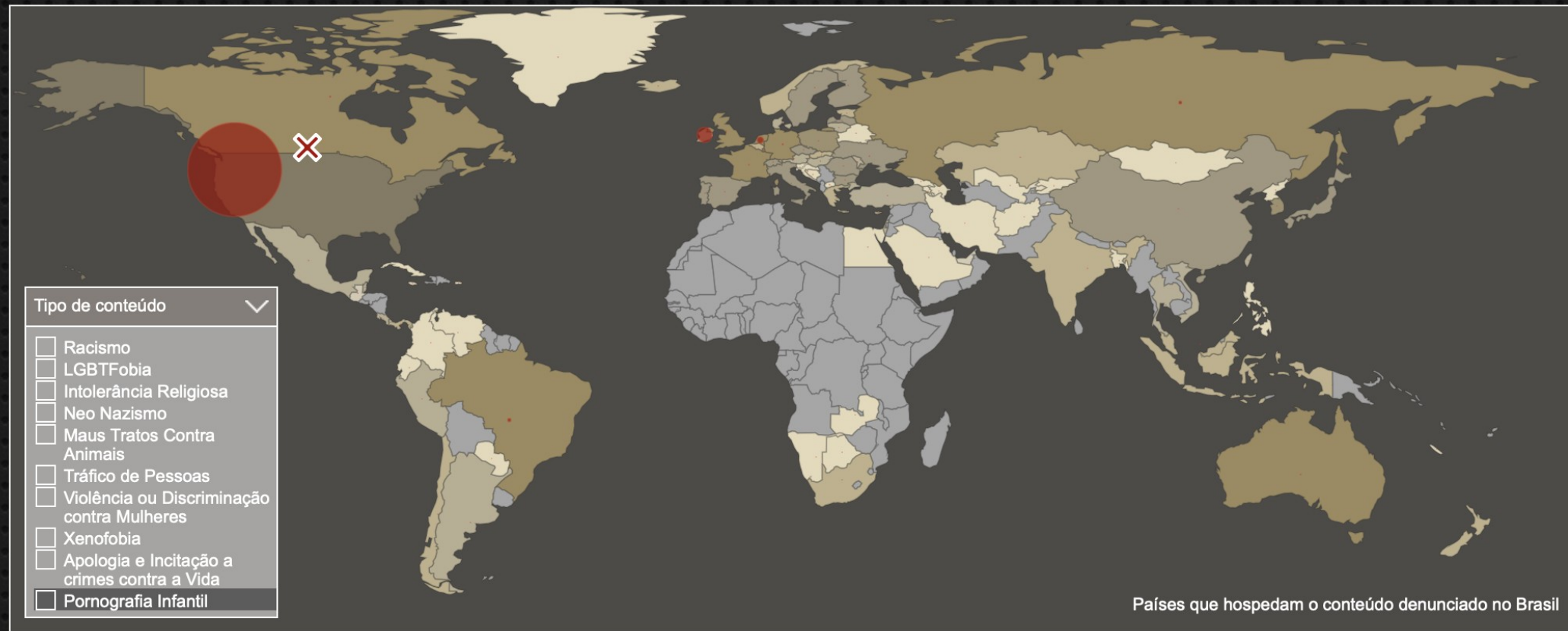
246.699

PÁGINAS REMOVIDAS

# Indicadores da Central Nacional de Denúncias de Crimes Cibernéticos

COMO UTILIZAR ESTE MAPA?

Em **15 anos**, a **Central de Denúncias** recebeu e processou **1.759.354** denúncias anônimas de **Pornografia Infantil** envolvendo **429.665** páginas (URLs) distintas (das quais **340.005** foram removidas) escritas em **10 idiomas** e hospedadas em **59.177** domínios diferentes, de **260** diferentes TLDs e conectados à Internet através de **64.921** números IPs distintos, atribuídos para **101** países em **6** continentes. As denúncias foram registradas pela população através dos **3** hotlines brasileiros que integram a Central Nacional de Denúncias de Crimes Cibernéticos. [Saiba mais sobre este projeto!](#)



ZOOM DO MAPA



ESCALA DA BOLHA



LINHA DO TEMPO



2006 a 2020

REALIZAÇÃO



PARCEIROS





# Operação Luz na Infância 2

Maior operação de combate à pornografia infantil da história no Brasil

## Dados da operação\*



**251**

pessoas presas



**579**

mandados de busca e apreensão



**2,6 mil**

policiais envolvidos

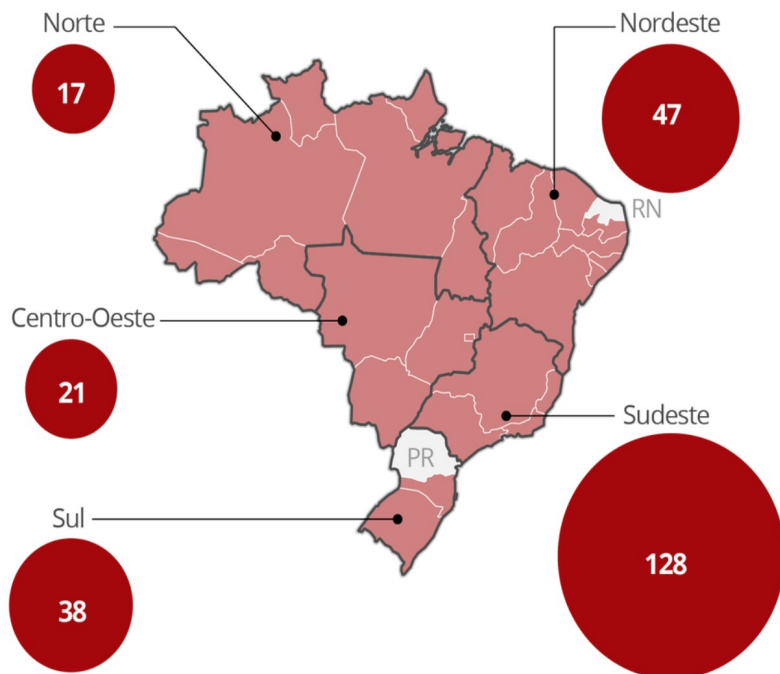


**1 milhão**

de arquivos analisados

## Prisões por região

■ Estados com operação



\*Até as 18h

Fonte: Ministério Extraordinário da Segurança Pública

Infográfico elaborado em: 17/05/2018

## PRISÕES EM FLAGRANTE

LUZ NA INFÂNCIA 4

ACRE	6
ALAGOAS	2
AMAZONAS	1
BAHIA	1
CEARÁ	1
DISTRITO FEDERAL	5
ESPÍRITO SANTO	5
GOIÁS	10
MARANHÃO	1
MATO GROSSO	5
MATO GROSSO DO SUL	4
MINAS GERAIS	10
PARÁ	3
PARAÍBA	1
PARANÁ	6
PERNAMBUCO	2
PIAUÍ	1
RIO DE JANEIRO	5
RIO GRANDE DO NORTE	1
RIO GRANDE DO SUL	4
RONDÔNIA	1
SANTA CATARINA	4
SÃO PAULO	61
SERGIPE	1
<b>TOTAL</b>	<b>141</b>



# Cooperação Multisetorial



Art. 227. É dever da família, da sociedade e do Estado assegurar à criança, ao adolescente e ao jovem, com absoluta prioridade, o direito à vida, à saúde, à alimentação, à educação, ao lazer, à profissionalização, à cultura, à dignidade, ao respeito, à liberdade e à convivência familiar e comunitária, além de colocá-los a salvo de toda forma de negligência, discriminação, exploração, violência, crueldade e opressão.

## São Paulo

[Página Inicial](#) > [Sala de Imprensa](#) > [Notícias](#) > MPF e SaferNet identificam mais de 6 mil sites de pornografia infantil

Pesquisar...



## Procuradoria da República em São Paulo

[Institucional](#)[Atuação](#)[Serviços](#)[Municípios](#)[PRDC](#)[Estágio Conosco](#)[Sala de Imprensa](#)[Editais e Administração](#)

## Notícias

CRIMINAL

5 DE SETEMBRO DE 2018 ÀS 14H55

## MPF e SaferNet identificam mais de 6 mil sites de pornografia infantil

Assessoria de Comunicação



Núcleo de Eventos

Denúncias feitas por internautas são objeto de 832 investigações em São Paulo

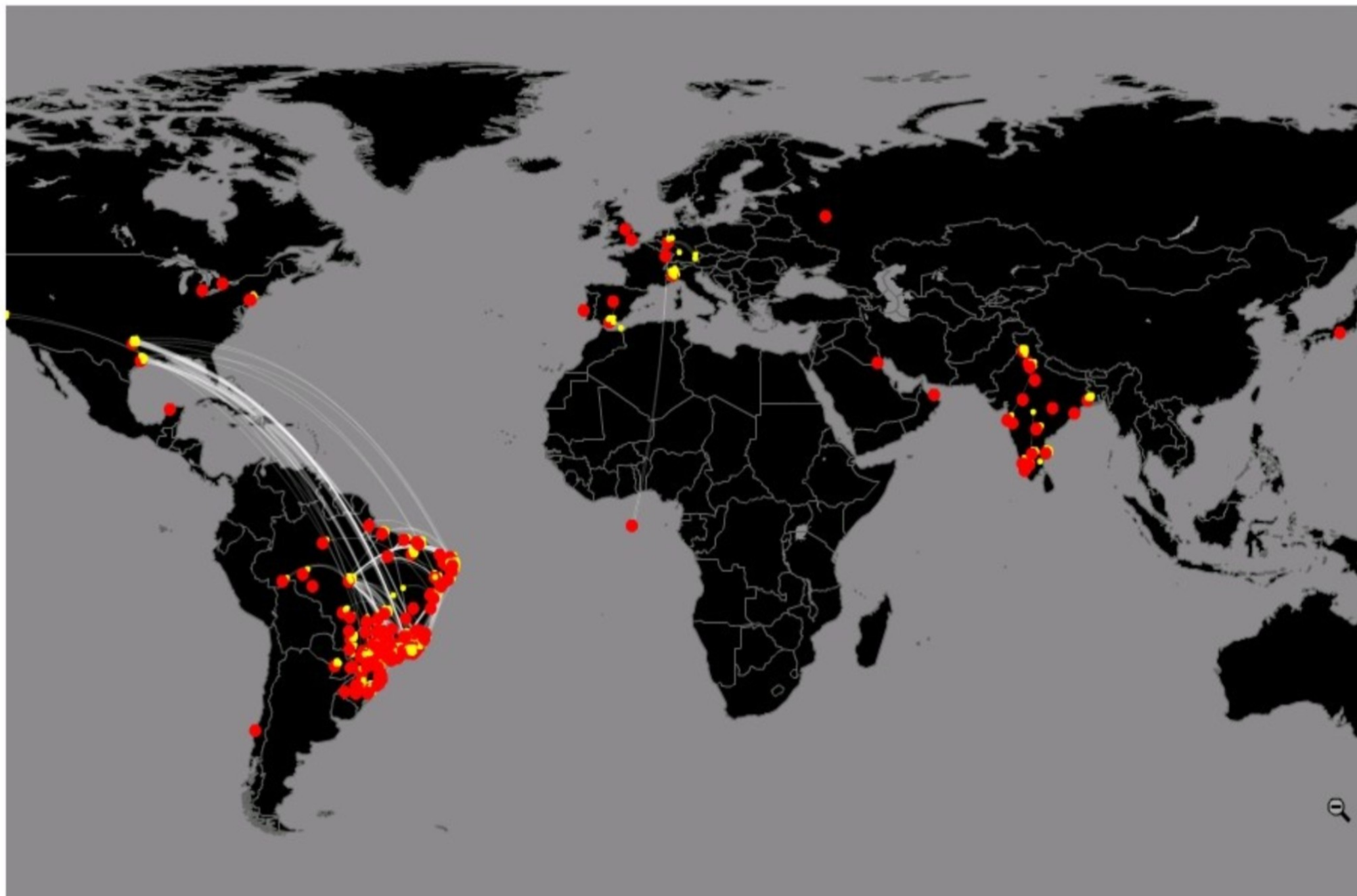


Imagem ilustrativa: Pixabay

O Ministério Público Federal e a ONG SaferNet Brasil identificaram mais de 6 mil sites com conteúdo criminoso, envolvendo principalmente abuso sexual e exploração de crianças e adolescentes. Os registros de pornografia infantil foram descobertos após denúncias feitas por internautas entre fevereiro de 2017 a agosto de 2018. Destas, 832 já são objeto de investigação pelo MPF. O combate a esse tipo de crime na internet se torna mais eficaz com o trabalho conjunto dessas duas instituições, formalizado por meio do Convênio Técnico e Operacional assinado em fevereiro do ano passado.

Graças à parceria, o MPF teve acesso à base de dados da SaferNet, que registrou mais de 57 mil denúncias no período analisado, reportadas por meio do site <http://www.denunciar.org.br>. Destas notificações, cerca de 6 mil continham informações que permitiram dar prosseguimento às investigações de crimes de pornografia infantil. Segundo o diretor e fundador da ONG, Thiago Tavares, o site recebe cerca de 100 novas denúncias diariamente.

Investigação 01 (iniciada em maio de 2008) – 1263 conexões em 12 países (874 no Brasil) – 300 agressores sexuais investigados



# Lei 11.829/08



## Presidência da República Casa Civil Subchefia para Assuntos Jurídicos

### LEI Nº 11.829, DE 25 DE NOVEMBRO DE 2008.

Altera a Lei nº 8.069, de 13 de julho de 1990 - Estatuto da Criança e do Adolescente, para aprimorar o combate à produção, venda e distribuição de pornografia infantil, bem como criminalizar a aquisição e a posse de tal material e outras condutas relacionadas à pedofilia na internet.

**O PRESIDENTE DA REPÚBLICA** Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei:

Art. 1º Os arts. 240 e 241 da [Lei nº 8.069, de 13 de julho de 1990](#), passam a vigorar com a seguinte redação:

“[Art. 240.](#) Produzir, reproduzir, dirigir, fotografar, filmar ou registrar, por qualquer meio, cena de sexo explícito ou pornográfica, envolvendo criança ou adolescente:

Pena – reclusão, de 4 (quatro) a 8 (oito) anos, e multa.

§ 1º Incorre nas mesmas penas quem agencia, facilita, recruta, coage, ou de qualquer modo intermedeia a participação de criança ou adolescente nas cenas referidas no caput deste artigo, ou ainda quem com esses contracenar.

§ 2º Aumenta-se a pena de 1/3 (um terço) se o agente comete o crime:

I – no exercício de cargo ou função pública ou a pretexto de exercê-la;

II – prevalecendo-se de relações domésticas, de coabitação ou de hospitalidade; ou

III – prevalecendo-se de relações de parentesco consanguíneo ou afim até o terceiro grau, ou por adoção, de tutor, curador, preceptor, empregador da vítima ou de quem, a qualquer outro título, tenha autoridade sobre ela, ou com seu consentimento.” (NR)

“[Art. 241.](#) Vender ou expor à venda fotografia, vídeo ou outro registro que contenha cena de sexo explícito ou pornográfica envolvendo criança ou adolescente:

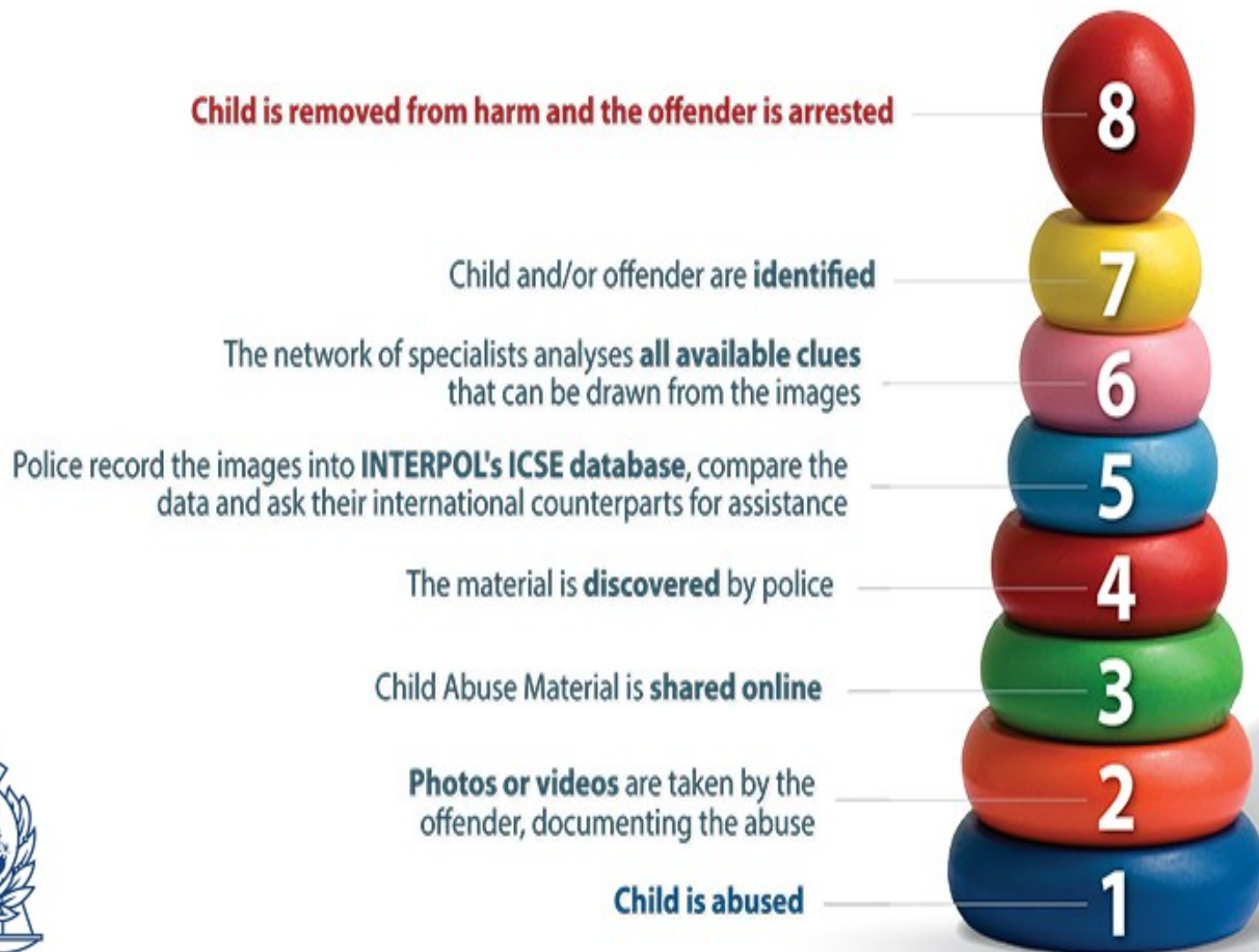
Pena – reclusão, de 4 (quatro) a 8 (oito) anos, e multa.” (NR)

Art. 2º A Lei nº 8.069, de 13 de julho de 1990, passa a vigorar acrescida dos seguintes arts. 241-A, 241-B, 241-C, 241-D e 241-E:

“[Art. 241-A.](#) Oferecer, trocar, disponibilizar, transmitir, distribuir, publicar ou divulgar por qualquer meio, inclusive por meio de sistema de informática ou telemático, fotografia, vídeo ou outro registro que contenha cena de sexo explícito ou pornográfica envolvendo criança ou adolescente:

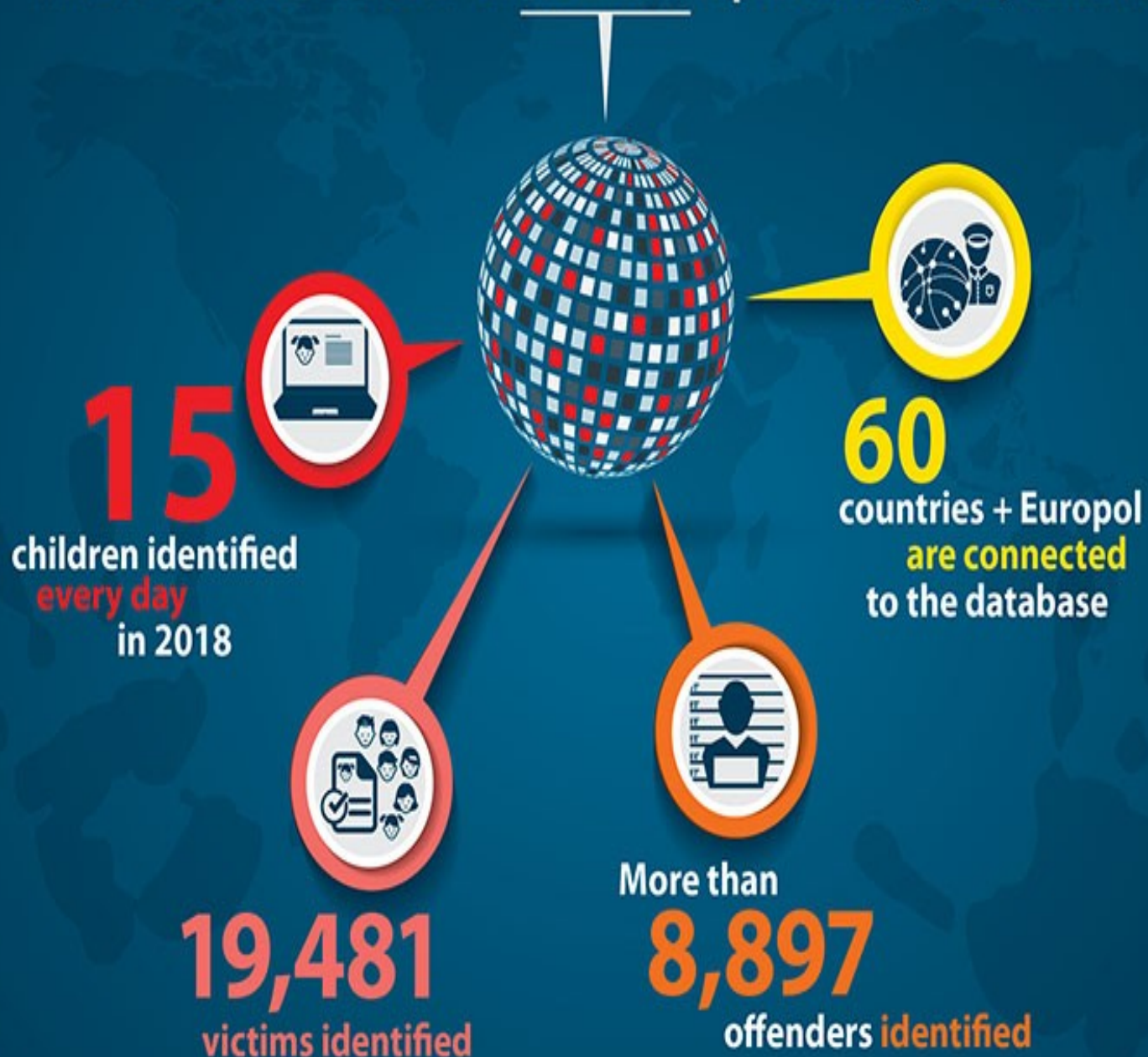
Pena – reclusão, de 3 (três) a 6 (seis) anos, e multa.

# 8 steps to identifying victims of child sexual abuse



INTERPOL

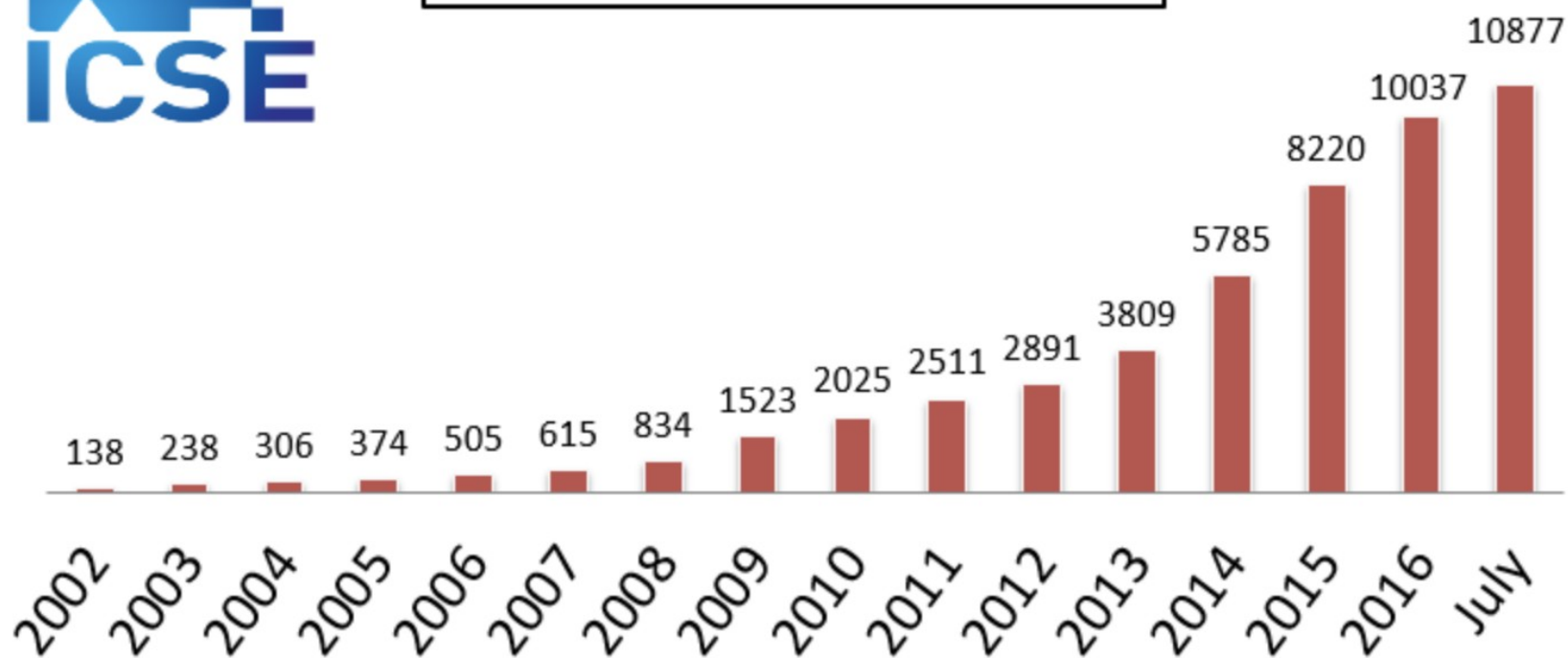
# INTERPOL's International Child Sexual Exploitation (ICSE) database



INTERPOL



## Identified victims in ICSE



The VGT Board of Management will next meet in November to address the issue of rising reports worldwide and work actively together with industry and other partners to prevent the spread of child abuse material and save children from victimization.

Fonte: <http://virtualglobaltaskforce.com/vgt-reports-of-child-exploitation-material-continue-to-increase/>

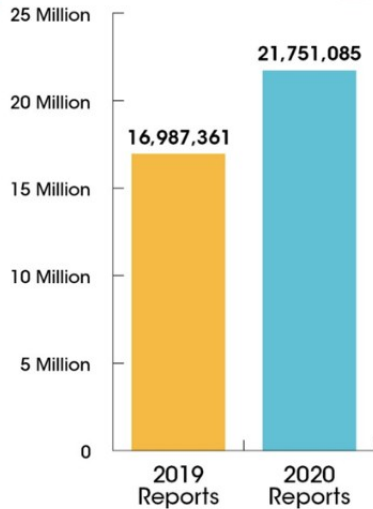
## Mandatory Reporting in the United States

18 U.S.C. § 2258A

- Stipulates U.S. based companies **shall** report instances of “apparent child pornography” to the CyberTipline
- Treats receipt of CyberTipline report as “preservation” request for 90 days
- Provides ESPs immunity for transfer of apparent child pornography images to the CyberTipline
- Specifies what the company **may** provide in each report
  - Suspect/uploader information
  - Historical information
  - Jurisdictional information

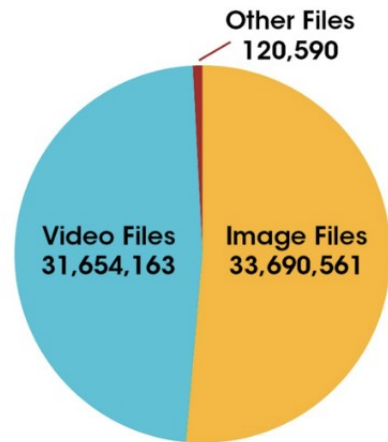


# By the Numbers



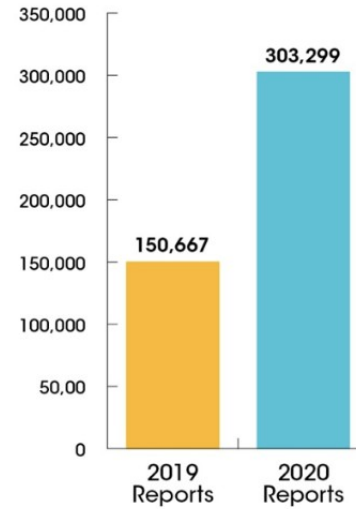
**Total Reports**

In 2020, reports to the CyberTipline increased by 28% from 2019.



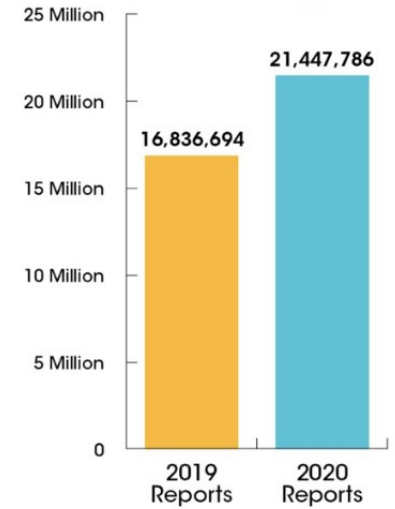
**Files Included in 2020 Reports**

The 21.7 million reports of child sexual exploitation made to the CyberTipline in 2020 included 65.4 million images, videos and other files. These materials contained suspected child sexual abuse material (CSAM) and other incident related content.



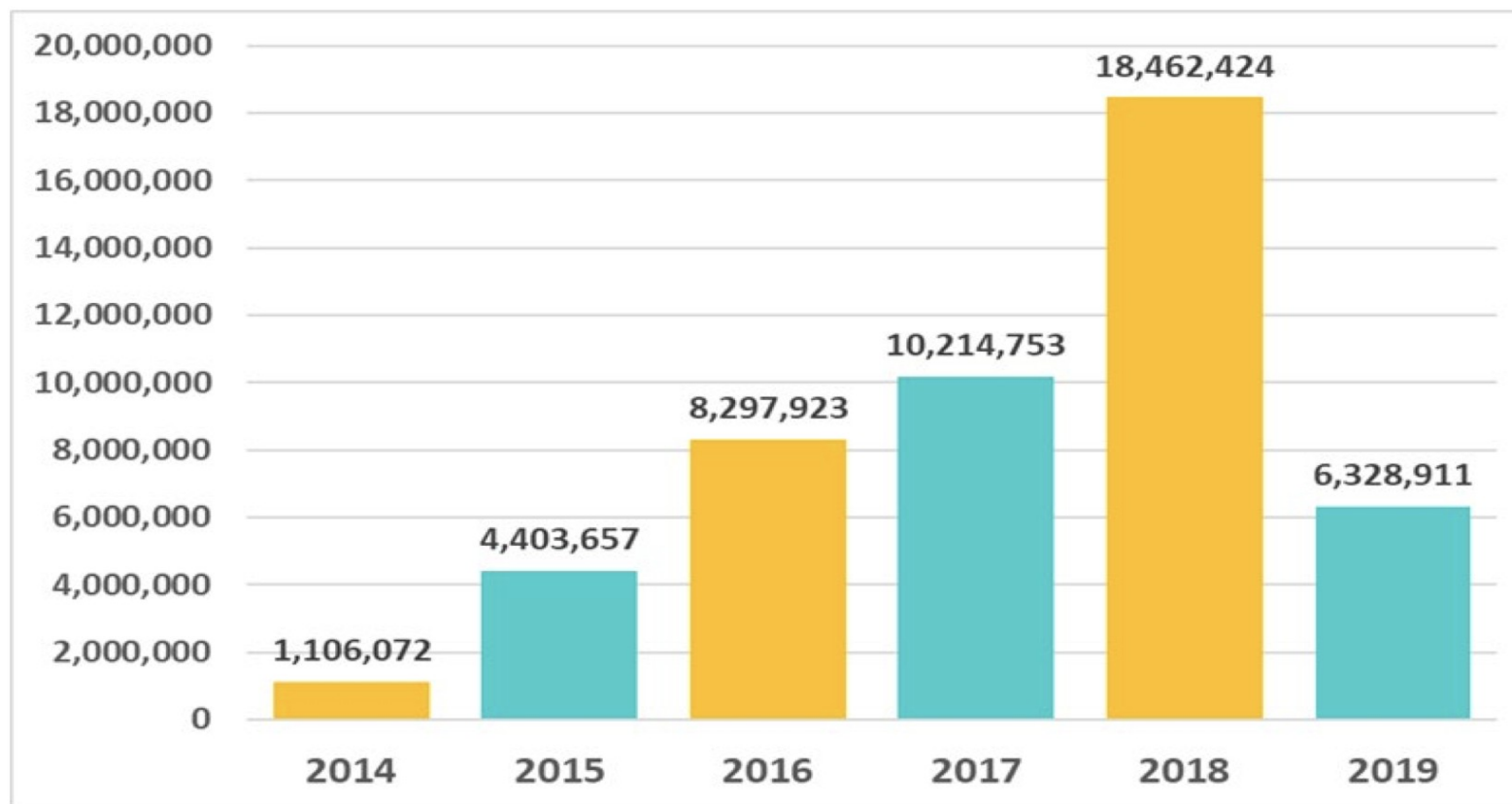
**Total Reports by Public**

Electronic Service Providers make the majority of reports, but reports of online sexual exploitation from the public more than doubled in 2020.



**Total Reports by ESP**

## II. Increase in CyberTipline Reports to NCMEC from 2014-2019<sup>5</sup>

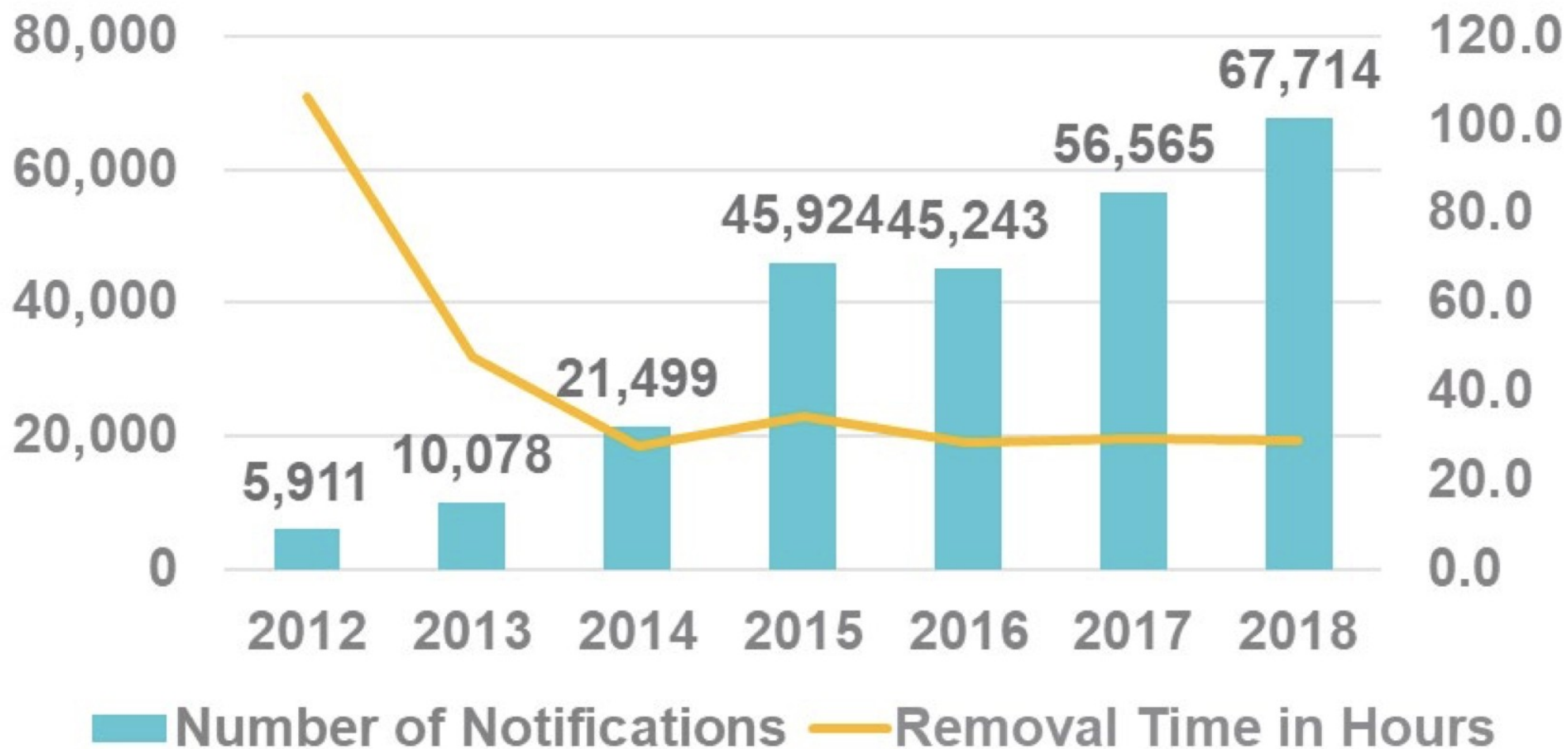


Fonte:  
<https://www.judiciary.senate.gov/imo/media/doc/Clark%20Testimony.pdf>

Multiple factors contribute to the exponential increase in reports to NCMEC's CyberTipline, including the following:

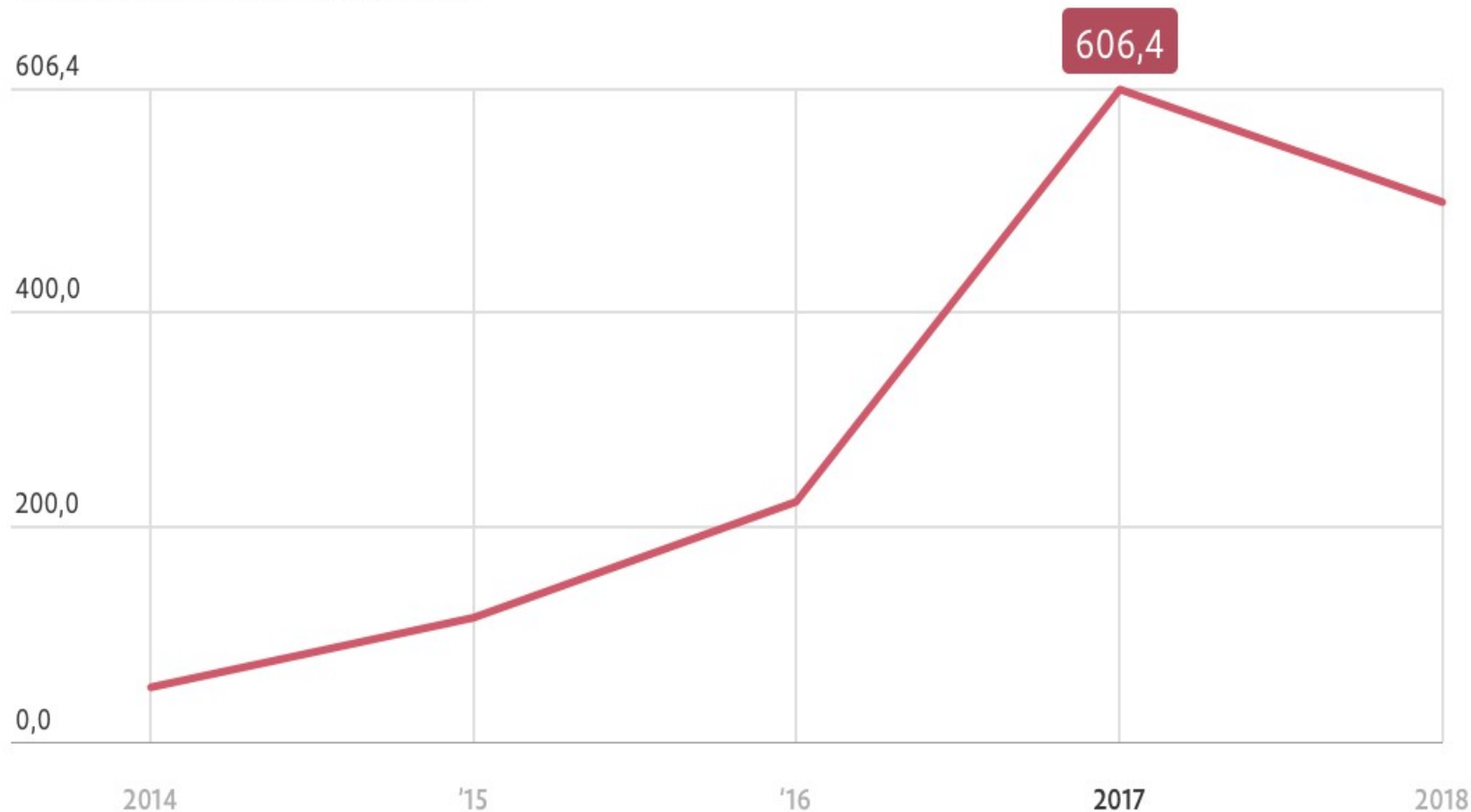
- Wide-spread voluntary adoption by ESPs of new technologies to locate and remove child sexual exploitation content from their platforms and services;
- Growing international scope of the crime;<sup>6</sup>
- Increased use of U.S.-based social media, mobile-based apps, and chat and photo-sharing programs by members of the public from around the world; and
- Decreased financial and access barriers to using the Internet to facilitate storing and sharing ever-larger volumes of child sexual abuse images and videos.

## Notices to Hosting Providers



Fonte: <https://www.judiciary.senate.gov/imo/media/doc/Clark%20Testimony.pdf>

Imagens de abuso sexual infantil compartilhadas a partir do Brasil, segundo relatórios internacionais



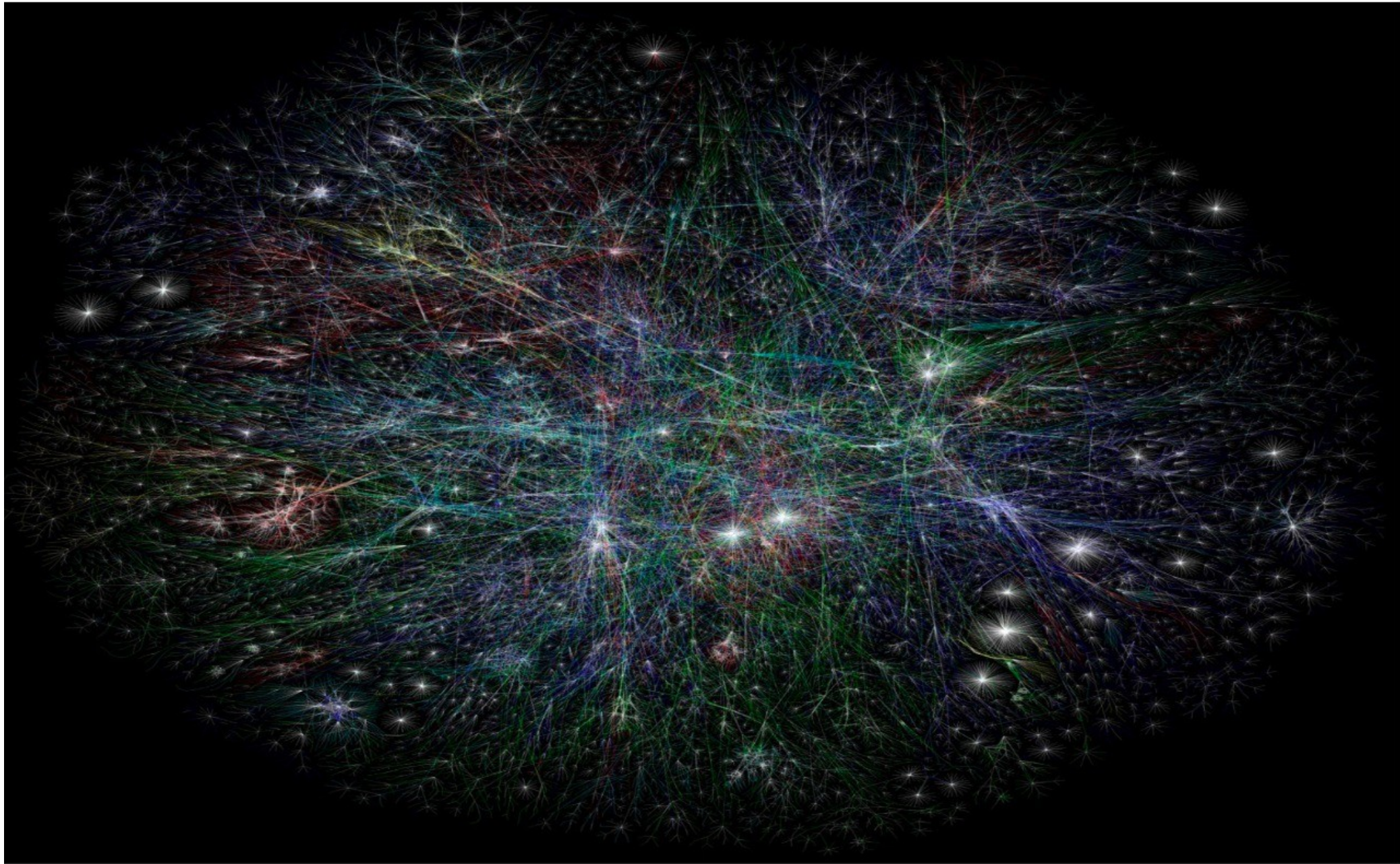
Aumento superior a **1.000%** em quatro anos

Fontes: Inhope 2017 e NCMEC, para casos em que vítima e agressor eram identificáveis

# Subsídios para avançarmos no debate sobre o PL 2630

- Premissas
- Moderação de Conteúdo em Escala
  - Modelos de negócio, fatores humanos e impacto social e econômico das escolhas regulatórias
  - Automação da Análise de Conteúdo em Escala
  - Indicadores de Qualidade usados (KPIs)
- Conclusão

Internet: ~70K Autonomous System (ASNs)



# THE THREE LAYERS OF DIGITAL GOVERNANCE

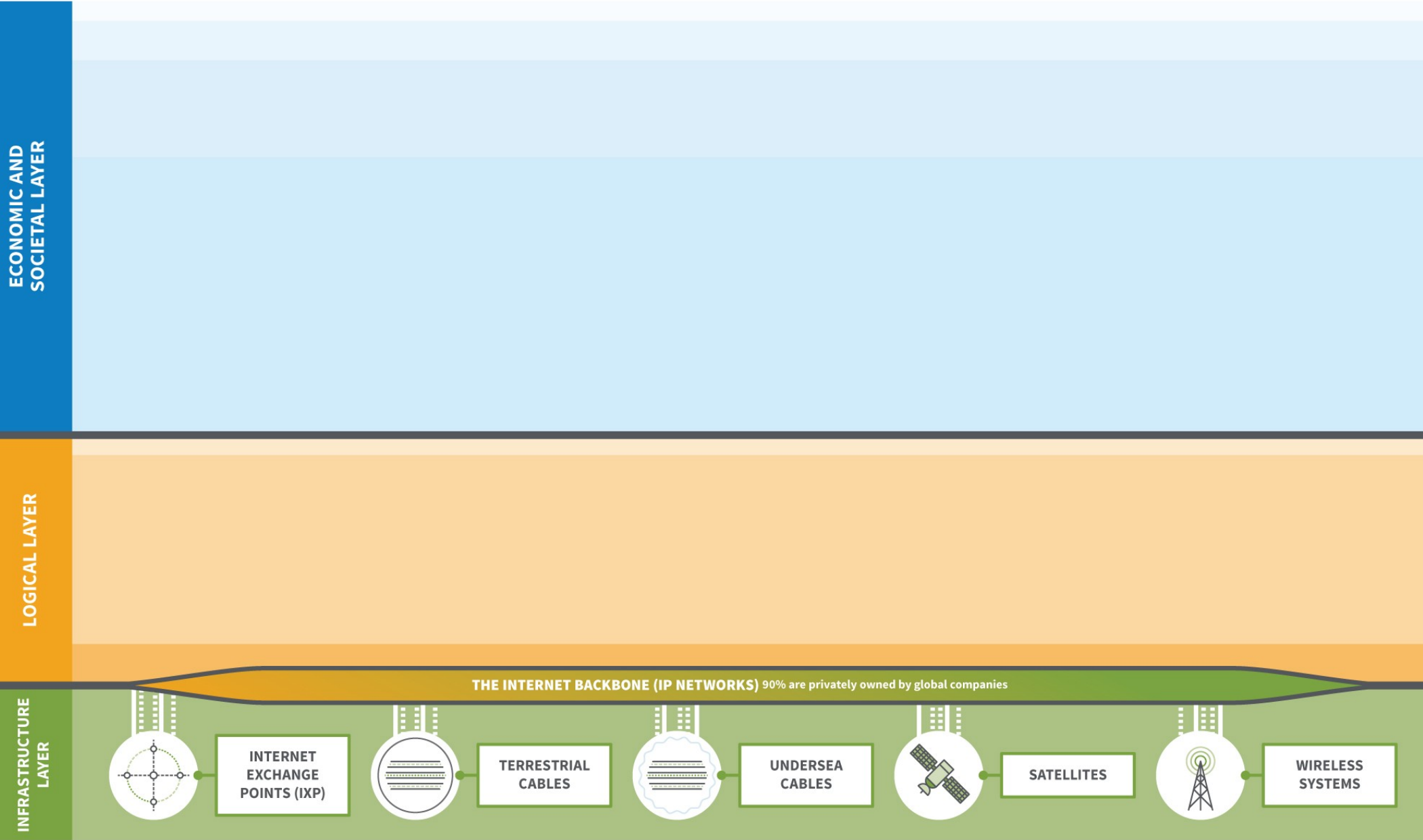


**ECONOMIC AND SOCIETAL LAYER**

**LOGICAL LAYER**

**INFRASTRUCTURE LAYER**

# THE INFRASTRUCTURE LAYER

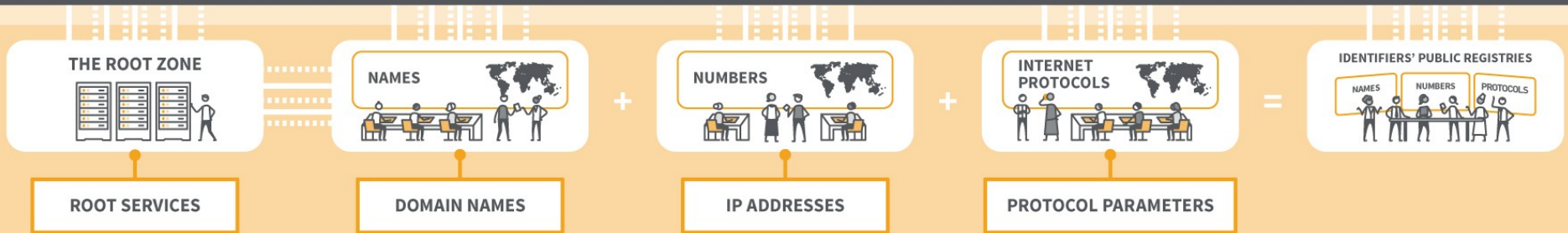




# THE LOGICAL LAYER

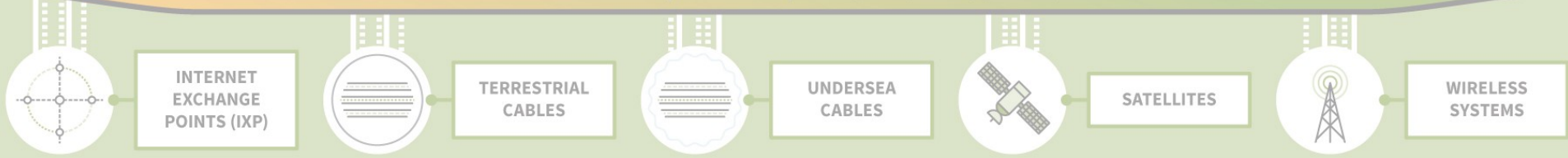
ECONOMIC AND SOCIETAL LAYER

LOGICAL LAYER



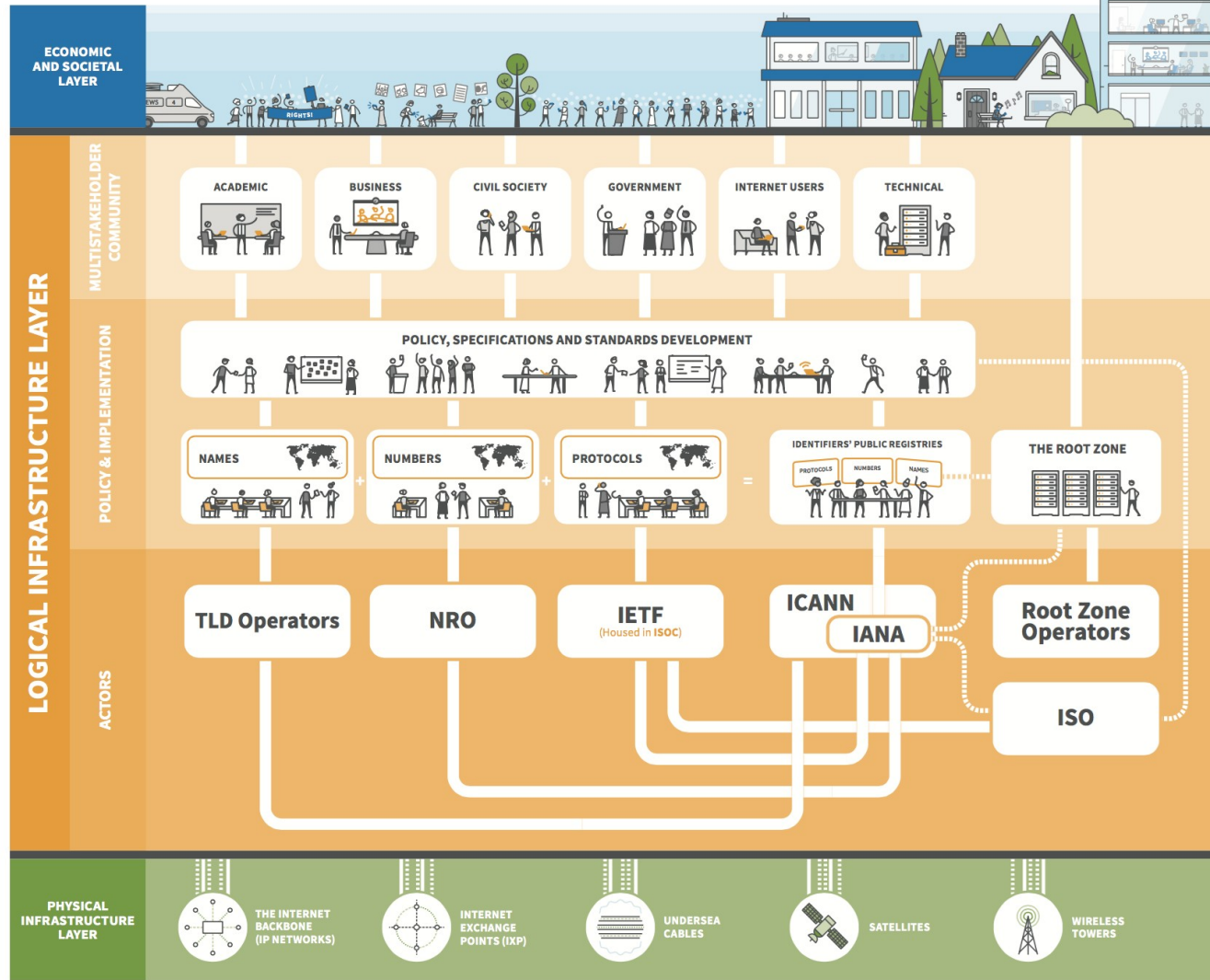
THE INTERNET BACKBONE (IP NETWORKS) 90% are privately owned by global companies

INFRASTRUCTURE LAYER



# WHO GOVERNS THE INTERNET'S LOGICAL INFRASTRUCTURE?

Layered on top of the Physical Infrastructure's thousands of networks and satellites, the Internet's Logical Infrastructure is what delivers One Internet for the world through Unique Identifiers (Names, Numbers, and Protocol Parameters). ICANN coordinates the administration of this layer in partnership with other technical communities to ensure the security, stability, resiliency, and integrity of this critical layer.



## TECHNICAL OPERATIONS

The technical Operating Community comprises multiple independent actors bound by common principles and mutual commitments that ensure its security and stability of the Logical Infrastructure of the Internet. Each actor's community develops policies and standards in an open, inclusive, and consensus-based approach.

### ACTORS

#### ICANN *Internet Corporation for Assigned Names and Numbers*

Helps coordinate the Internet's systems of unique identifiers including domain names and IP addresses, as well as manages the IETF's protocol parameters.

**IANA**, the Internet Assigned Numbers Authority, is a function housed and operated within ICANN. It acts as the top-level allocator for blocks of IP addresses and AS numbers, proposes creation of and changes to DNS top-level domains, and manages lists of unique identifiers used in Internet protocols.  
[www.icann.org](http://www.icann.org)  
[www.iana.org](http://www.iana.org)

#### IETF *Internet Engineering Task Force*

Develops and promotes a wide range of Internet standards dealing in particular with standards of the Internet protocol suite. Their technical documents influence the way people design, use, and manage the Internet. The IETF operates under the Internet Society (ISOC) with architectural oversight provided by the Internet Architecture Board (IAB).  
[www.ietf.org](http://www.ietf.org)

#### ISO *International Organization for Standardization*

Standardizes, among many other things, the official names and postal codes of countries, dependent territories, special areas of geographic significance.  
[www.iso.org](http://www.iso.org)

#### NRO *Number Resource Organization*

A coordinating body for the five Regional Internet Registries (RIRs). The RIRs manage the distribution of IP addresses and Autonomous System Numbers in their regions of the world.  
[www.nro.net](http://www.nro.net)  
 AFRNIC [www.afrinic.net](http://www.afrinic.net)  
 APNIC [www.apnic.net](http://www.apnic.net)  
 ARIN [www.arin.net](http://www.arin.net)  
 LACNIC [www.lacnic.net](http://www.lacnic.net)  
 RIPE NCC [www.ripe.net](http://www.ripe.net)

#### TLD Operators *Top Level Domain Operators*

Organizations responsible for the management of the Top Level Domains such as: Generic TLDs (.com, .biz, .edu), Country Code TLDs (.fr, .us, .cn) operators, and Internationalized Country Code for non-latin alphabet systems (Chinese, Arabic)—among others.  
[www.wikipedia.org/wiki/Top-level\\_domain](http://www.wikipedia.org/wiki/Top-level_domain)

#### Root Zone Operators

12 independent organisations operate the 13 authoritative name servers (A through M) that serve the Domain Name System (DNS) root zone. The name servers are a network of hundreds of physical servers located in many countries around the world.  
[www.root-servers.org](http://www.root-servers.org)

### MULTISTAKEHOLDER COMMUNITY

#### Academic

- Institutions of higher learning
- Academic thought leaders
- Professors & students

#### Business

- Private-sector companies from across industries
- Industry and trade associations

#### Civil Society

- International organizations
- Non-governmental organizations
- Non-profit organizations
- Think Tanks

#### Government

- National governments
- Distinct economies recognized in international fora
- Multinational governmental and treaty organizations
- Public authorities (with a direct interest in global Internet Governance)

#### Internet Users

- Private citizens interested in regional or global Internet Governance

#### Technical

- Internet engineers
- Computer engineers
- Software developers
- Network operators

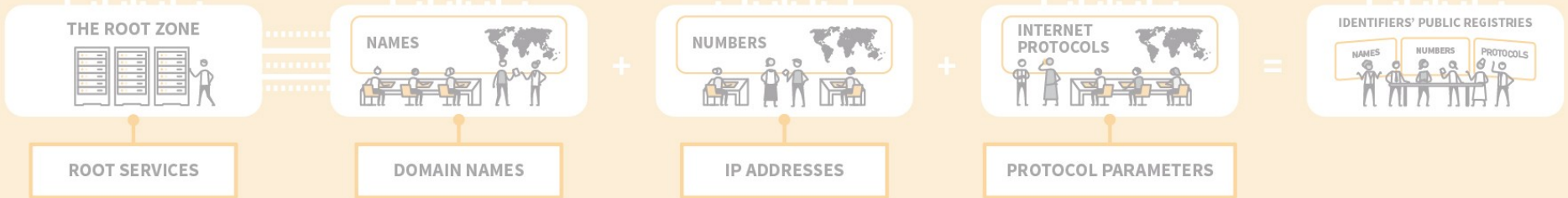
DRAFT

# THE ECONOMIC AND SOCIETAL LAYER

ECONOMIC AND SOCIETAL LAYER

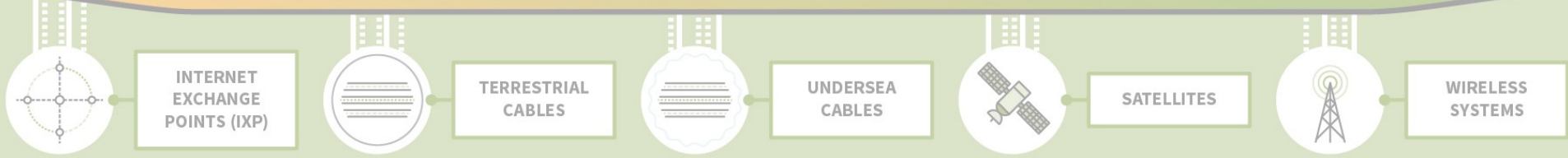


LOGICAL LAYER



THE INTERNET BACKBONE (IP NETWORKS) 90% are privately owned by global companies

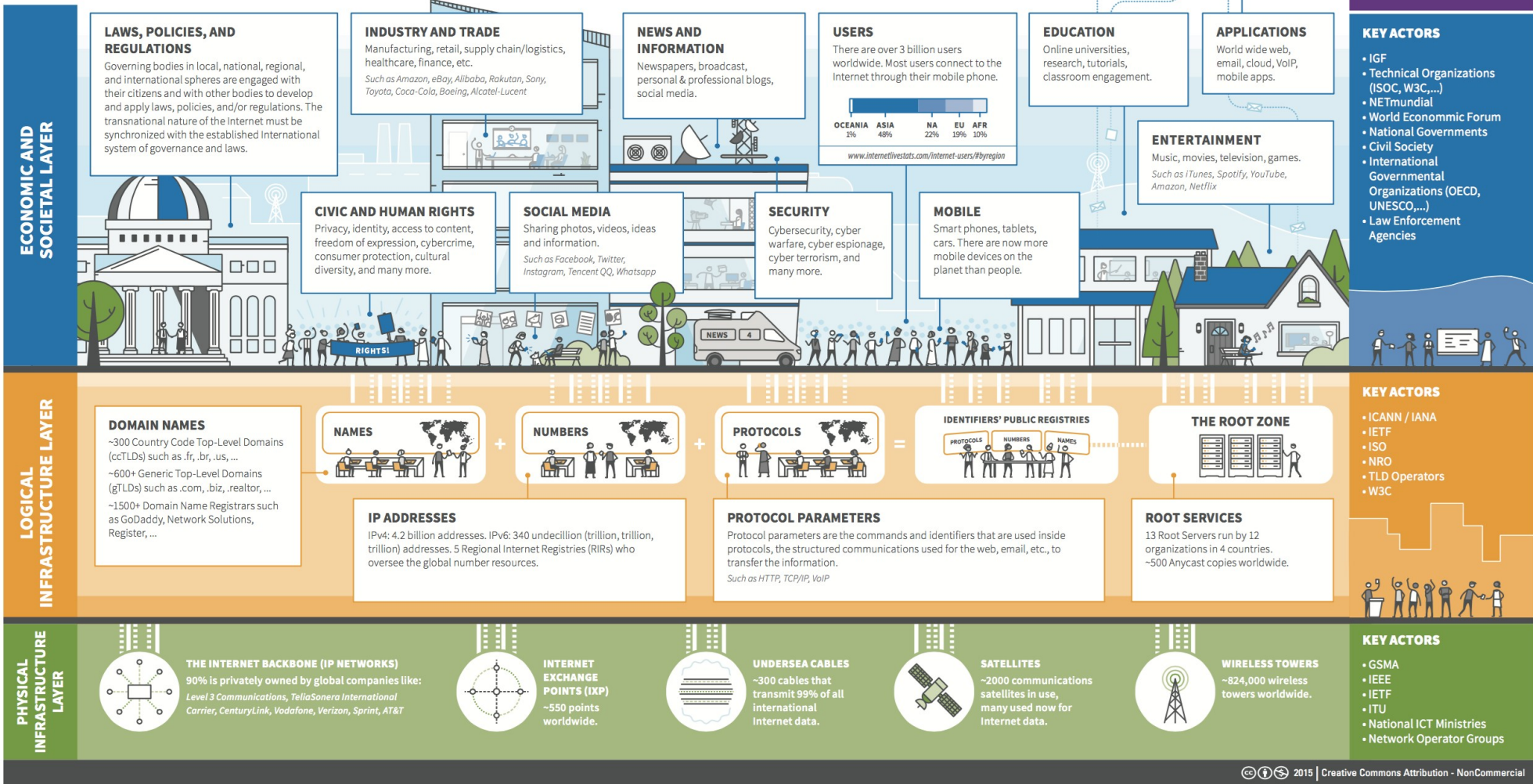
INFRASTRUCTURE LAYER



# THE THREE LAYERS OF DIGITAL GOVERNANCE

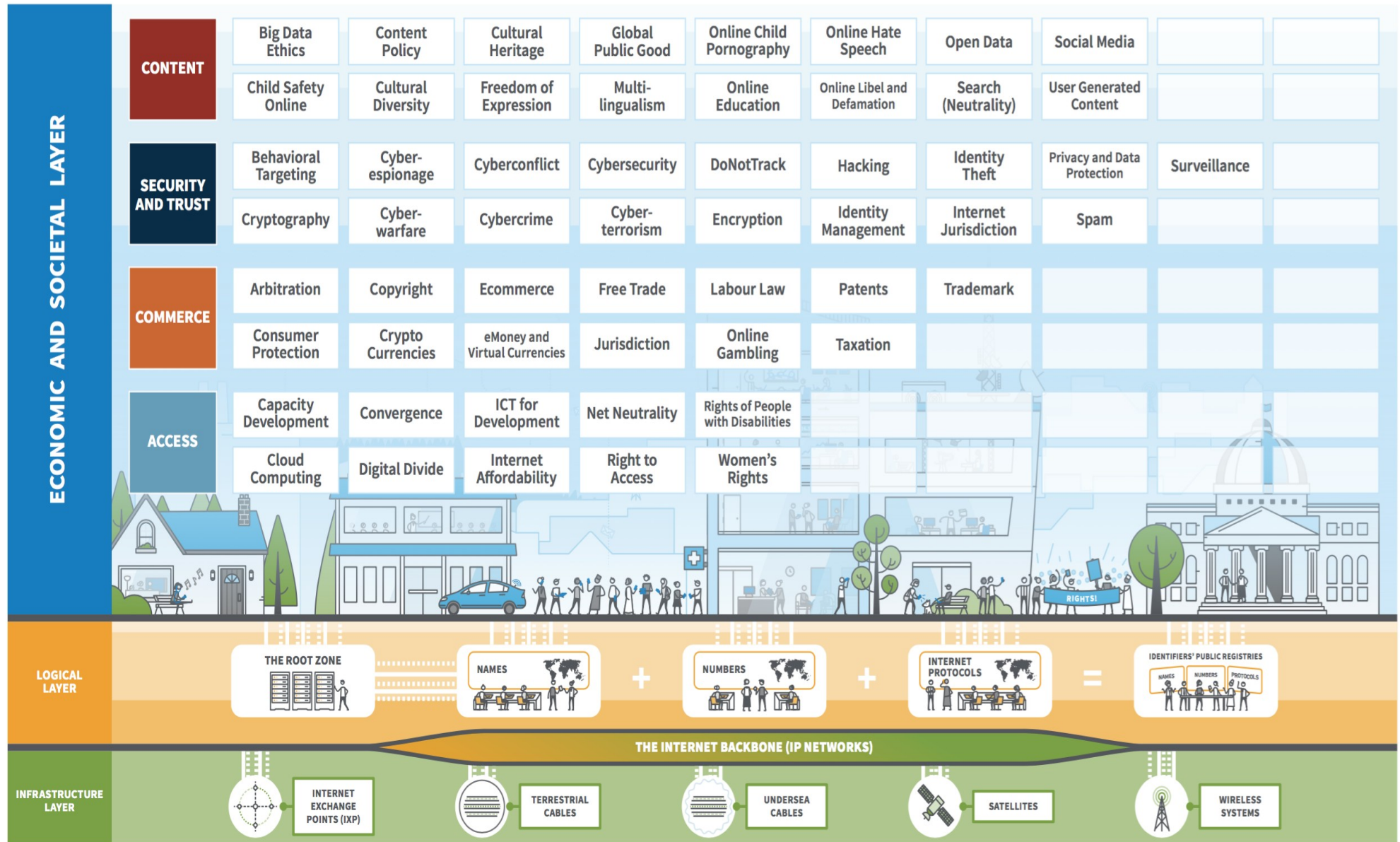
No one person, government, organization, or company governs the digital infrastructure, economy, or society. Digital governance is achieved through the collaborations of Multistakeholder experts acting through polycentric communities, institutions, and platforms across national, regional, and global spheres. Such Digital Governance is stratified into three layers to address infrastructure, economic, and societal issues with solutions. For a map of Digital Governance Issues and Solutions across all three layers, visit <https://map.netmundial.org>

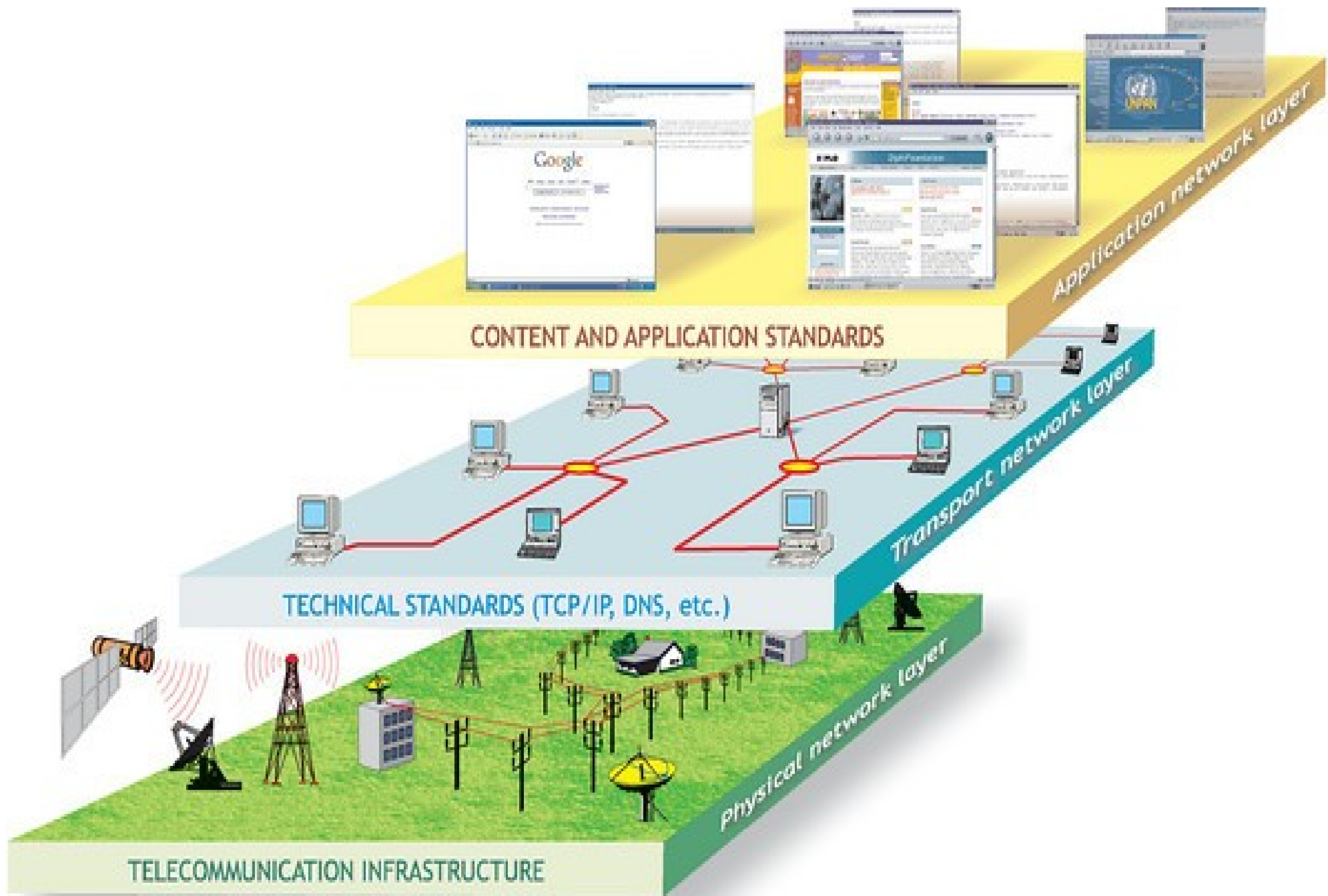
**MULTISTAKEHOLDER COLLABORATIONS**  
Solutions to issues in each layer include policies, best practices, standards, and specifications developed by the collaborations of expert stakeholders from actors in business, government, academia, technical, and civil society.



# THE ECONOMIC AND SOCIETAL LAYER OF DIGITAL GOVERNANCE

No one institution is able to design, develop, and implement solutions for the many Economic and Societal issues. Solutions to these issues require distributed, innovative, and collaborative issue-specific networks, coalescing organizations, experts, and stakeholders from governments, international organizations, the private sector, the technical community, and civil society. Solutions include policies, standards, specifications, best practices, and tools.





2018



C O N T E N T M O D E R A T I O N

AT SCALE

<http://comoatscale.com>

# Moderação de Conteúdo em Escala

Modelos de negócios, fatores humanos e impactos sociais e econômicos das escolhas regulatórias



## Content Moderation Solutions Market Segmentation

### Component

- Software/Tools/Platforms
  - *On-premise*
  - *Cloud*
- Services
  - *Professional Services*
  - *Managed Services*

### Enterprise Size

- Small and Medium Enterprises
- Large Enterprises

### Region

- North America
- Europe
- Asia Pacific
- Middle East & Africa
- South America

### Industry

- Media & Entertainment
- Retail & Ecommerce
- Packaging & Labelling
- Healthcare & Life-sciences
- Automotive
- Government
- Telecom
- Others (BFSI, Energy & Utilities)

# Content Moderation Solutions Market: Key Vendor Strategies

Combination of Human & Artificial Intelligence

Higher Penetration in E-Commerce

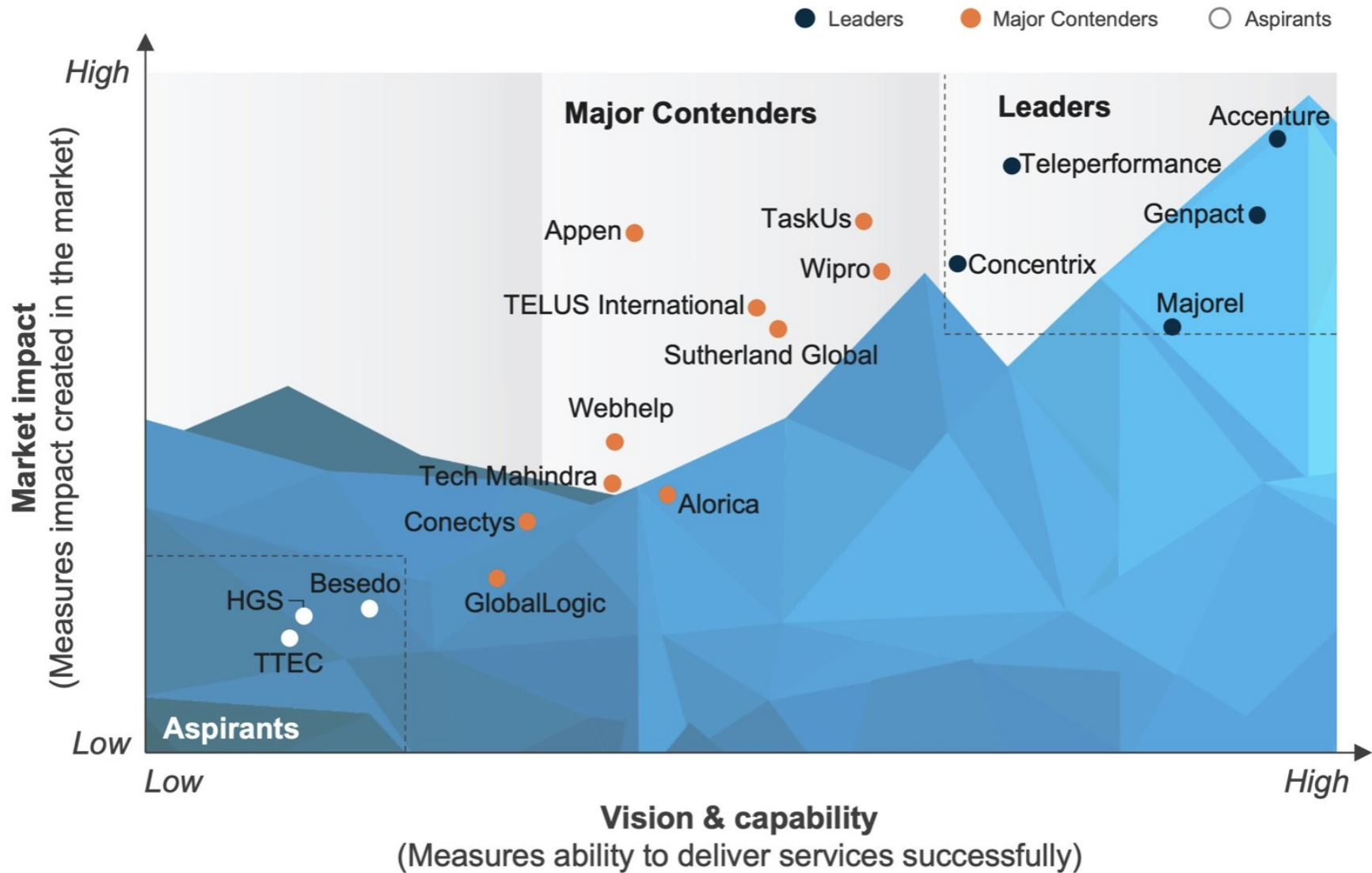
Upholding Freedom of Speech

Adopting Employee-centric Policies

Real-time Content Moderation



# Everest Group Trust and Safety – Content Moderation Services PEAK Matrix® Assessment 2021<sup>1</sup>



<sup>1</sup> Assessments for Appen and GlobalLogic exclude service provider inputs and are based on Everest Group's proprietary Transaction Intelligence (TI) database, service provider public disclosures, and Everest Group's interactions with trust and safety – content moderation services clients

Source: Everest Group (2021)

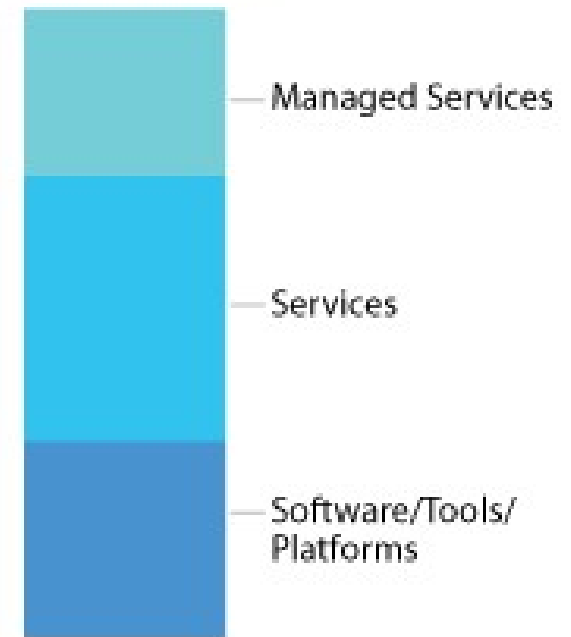
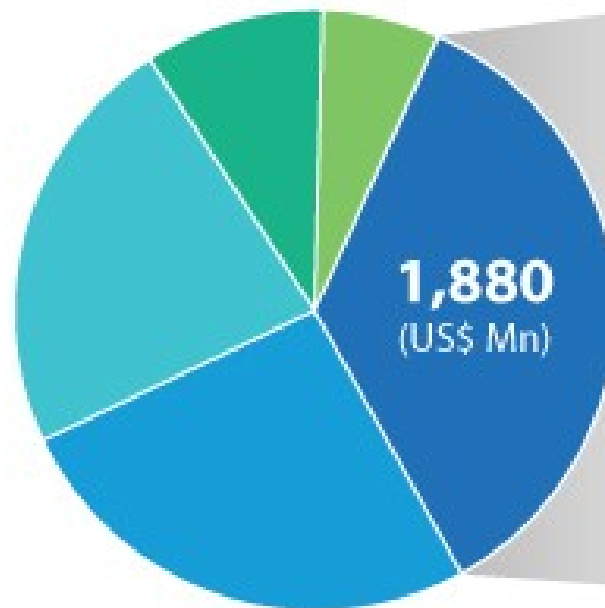
# Content Moderation Solutions Market: North America Analysis

Market Size by Region, 2019

Market Size by Component, 2019

North America

- North America
- Europe
- Asia Pacific
- Middle East & Africa
- South America

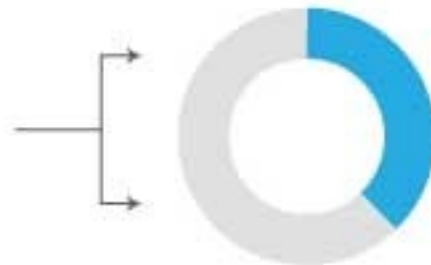


# Content Moderation Solutions Market

## Regional Analysis



North America



Europe



East Asia



CAGR  
(2020-2030)

Absolute \$ Opportunity  
(2020-2030)

**Note:**

Pie Chart indicates market share by Region

Arrow indicates the relative growth of the market in the region

\$ gradient fill represents absolute \$ opportunity created in respective region

Source: Fact.MR

Fact.MR

<https://www.factmr.com/report/4522/content-moderation-solutions-market>

REPORT

# Global Content Moderation Solutions Market Report and Forecast 2021-2026

213 pages

2021

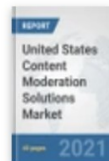
## RELATED PRODUCTS



### Global Content Moderation Solutions Market Report and Forecast 2021-2026

REPORT | 213 PAGES | JULY 2021 | REGION: GLOBAL

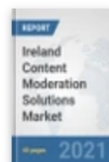
€ 2402



### United States Content Moderation Solutions Market: Prospects, Trends Analysis, Market Size and Forecasts up to 2027

REPORT | 40 PAGES | AUGUST 2021 | REGION: UNITED STATES

€ 1780



### Ireland Content Moderation Solutions Market: Prospects, Trends Analysis, Market Size and Forecasts up to 2027

REPORT | 40 PAGES | AUGUST 2021 | REGION: IRELAND

€ 1780

### 13. South America Content Moderation Solution Market Analysis and Forecast

#### 13.1. Key Findings

#### 13.2. Impact Analysis of Drivers and Restraints

#### 13.3. Content Moderation Solution Market Size (US\$ Mn) Forecast, by Components, 2017 - 2027

##### 13.3.1. Software/Tools/Platforms

###### 13.3.1.1. Cloud

###### 13.3.1.2. On-premise

##### 13.3.2. Services

###### 13.3.2.1. Professional Services

###### 13.3.2.2. Managed Services

#### 13.4. Content Moderation Solution Market Size (US\$ Mn) Forecast, by Enterprise Size, 2017 - 2027

##### 13.4.1. Small and Medium Enterprises

##### 13.4.2. Large Enterprises

#### 13.5. Content Moderation Solution Market Size (US\$ Mn) Forecast, by Industry, 2017 - 2027

##### 13.5.1. Media & Entertainment

##### 13.5.2. Retail &E-commerce

##### 13.5.3. Packaging &Labeling

##### 13.5.4. Healthcare & Life Sciences

##### 13.5.5. Automotive

##### 13.5.6. Government

##### 13.5.7. Telecom

##### 13.5.8- Others (BFSI, Energy & Utilities)

#### 13.6. Content Moderation Solution Market Size (US\$ Mn) Forecast, by Country & Sub-region, 2017 - 2027

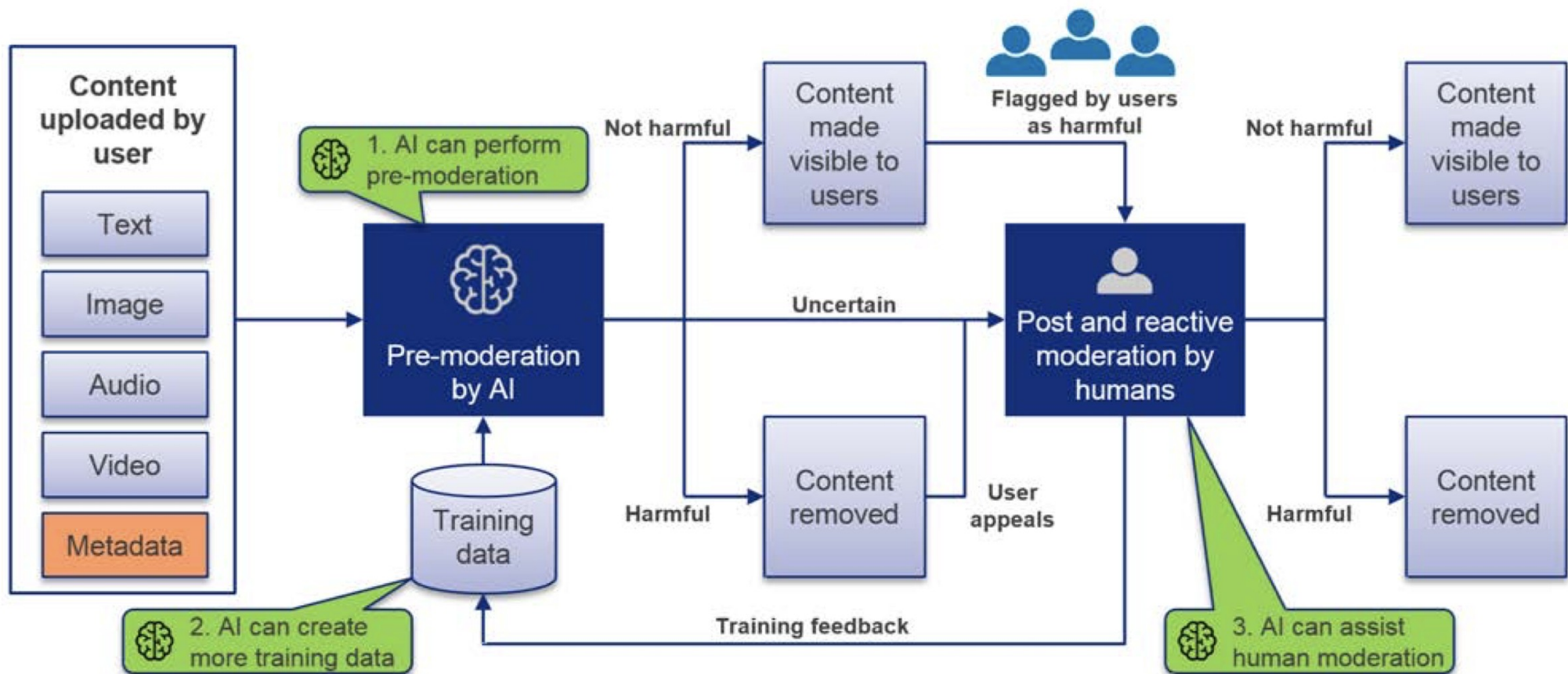
##### 13.6.1. [Brazil](#)

##### 13.6.2. Rest of South America

<https://www.researchandmarkets.com/reports/5401679/global-content-moderation-solutions-market-report#relc0-4851929>

# Moderação de Conteúdo em Escala

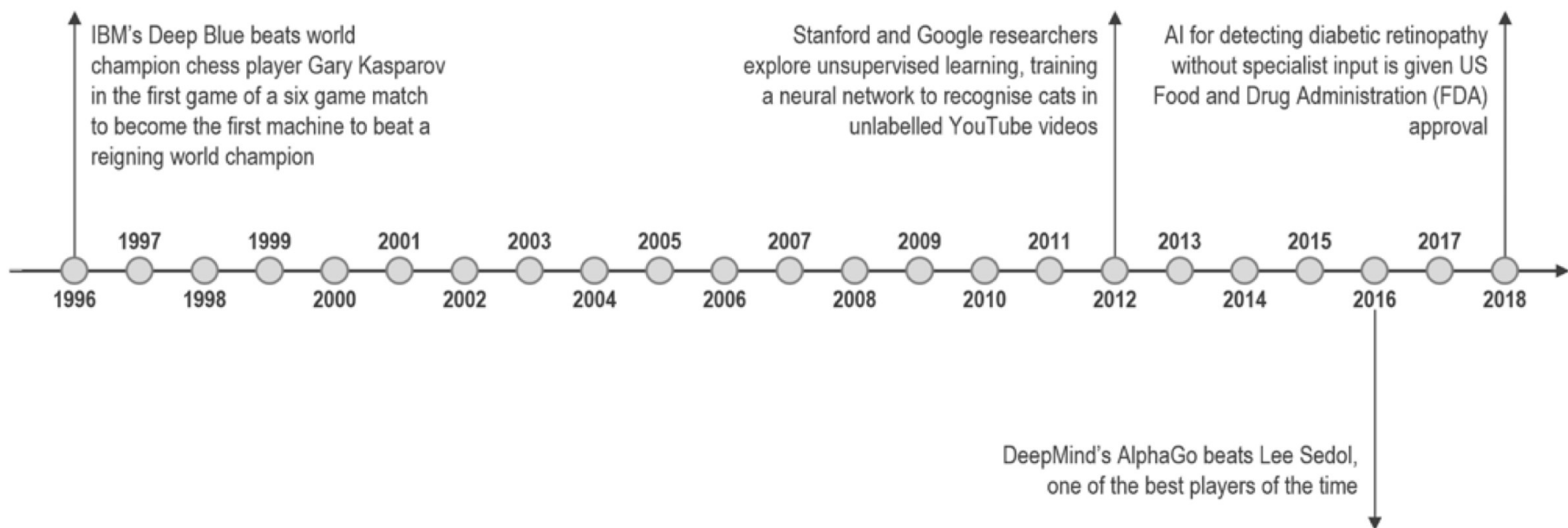
Automação da Análise de Conteúdo em Escala



**Figure 1** – There are three key ways in which AI can improve the effectiveness of the typical online content moderation workflow (SOURCE: Cambridge Consultants)

Source: [https://www.ofcom.org.uk/\\_\\_data/assets/pdf\\_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf](https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf)

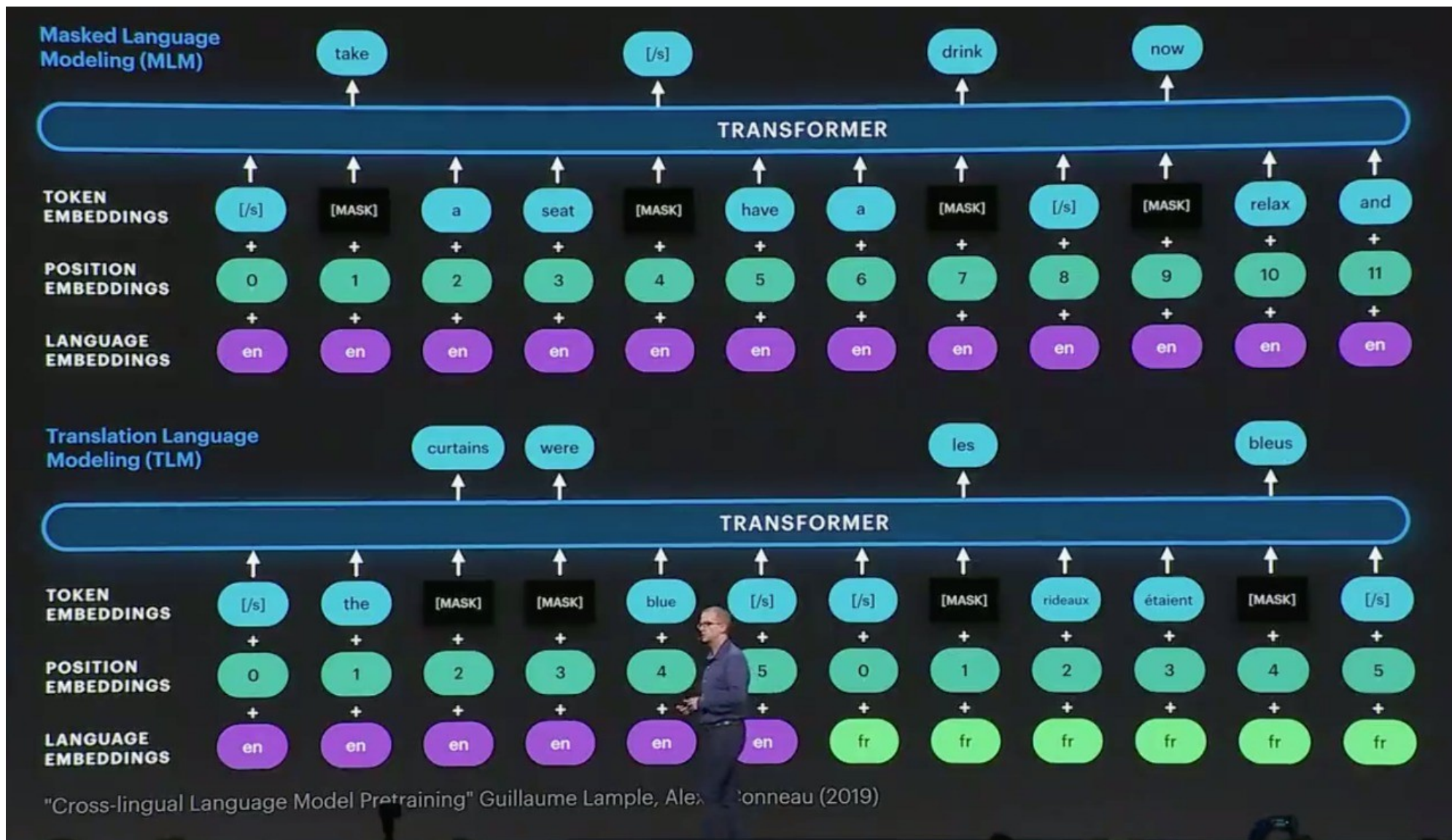




**Figure 6** – Key milestones in AI capabilities have been reached at an increasing rate (**SOURCE:** Cambridge Consultants)

Source: [https://www.ofcom.org.uk/\\_\\_data/assets/pdf\\_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf](https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf)

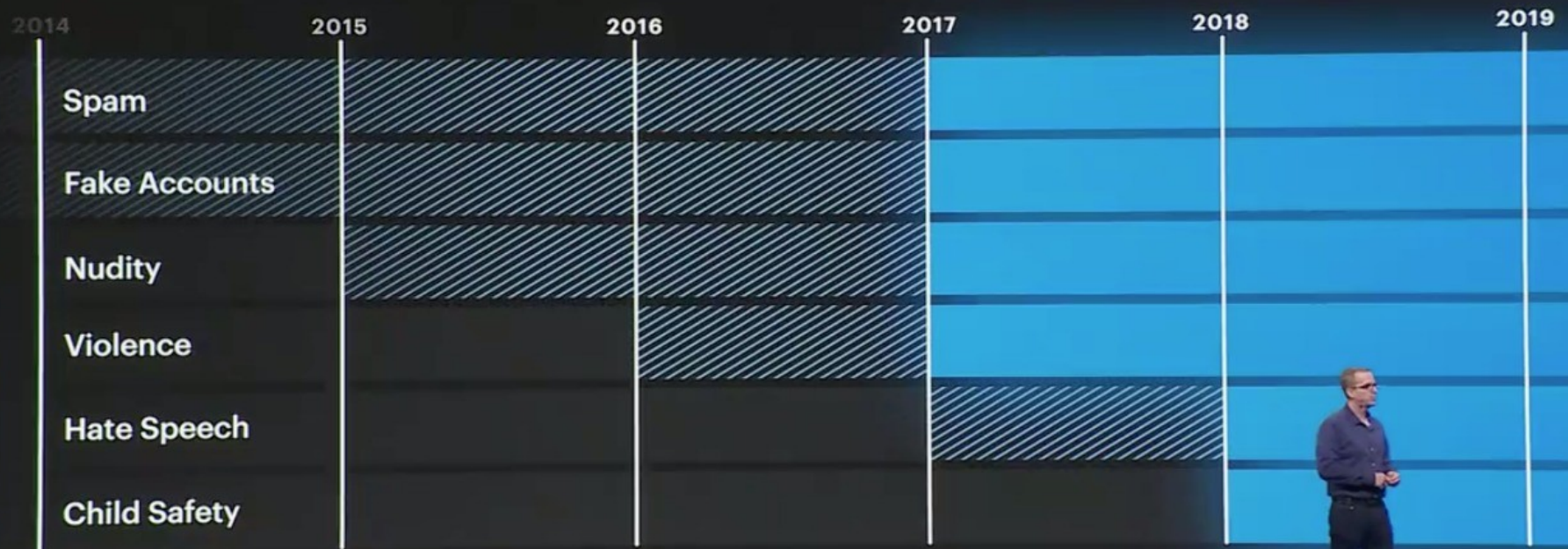




Fonte: Mike Schroepfer, F8 2019:  
<https://developers.facebook.com/videos/f8-2019/day-2-keynote/>

# Recent AI Advances

STARTED PRIMARY

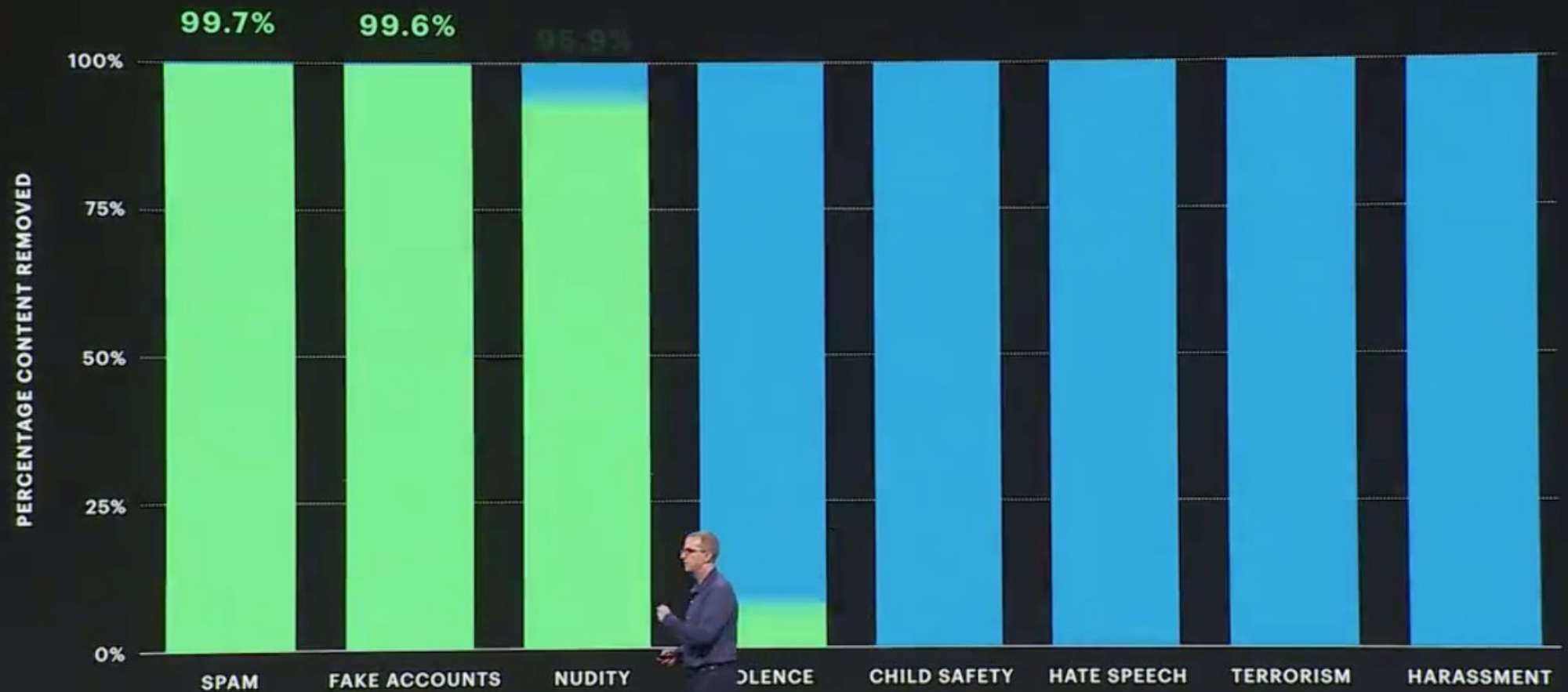


incluindo o conteúdo reportado por pessoas.

Fonte: Mike Schroepfer, F8 2019:

<https://developers.facebook.com/videos/f8-2019/day-2-keynote/>

# Violating Content Actioned in Q3 2018 Before People on Facebook Reported it



Facebook Transparency Report

Fonte: Mike Schroepfer, F8 2019:

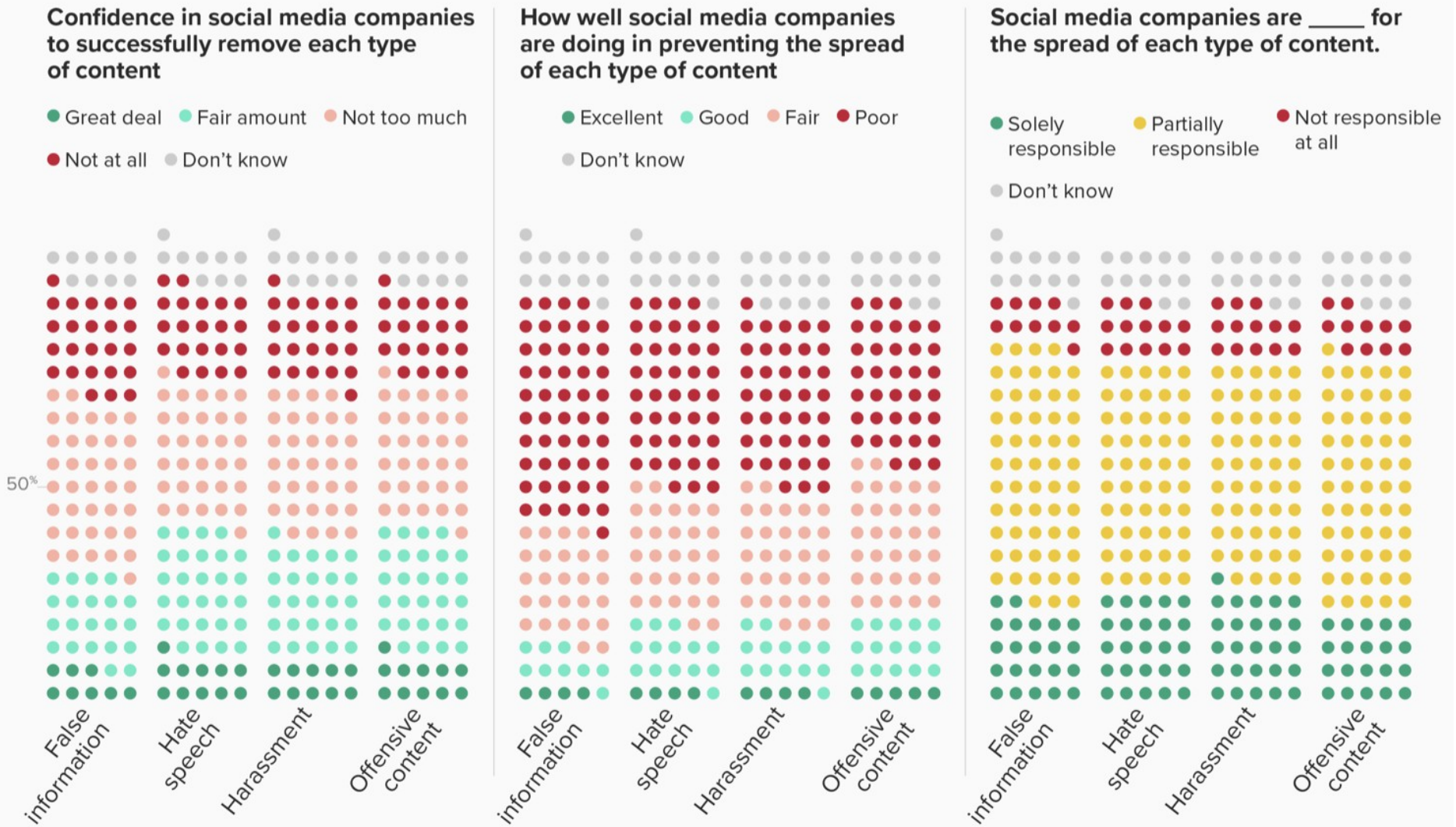
<https://developers.facebook.com/videos/f8-2019/day-2-keynote/>

# Content Moderation At Scale

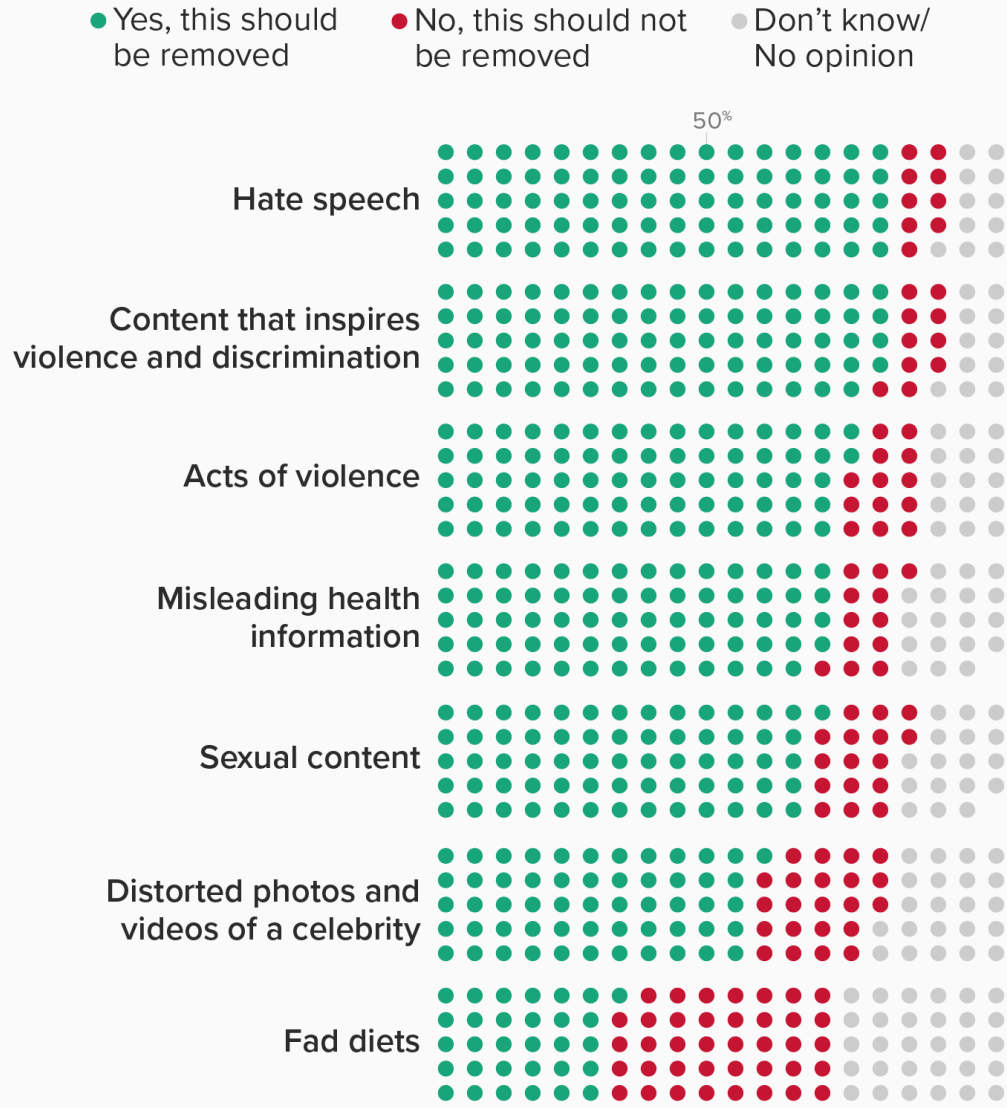


# Users Take Dim View of Social Media's Efforts to Moderate Content

Many have little to no confidence in companies' ability to handle harmful posts



# The Types of Content Users Find Most Objectionable

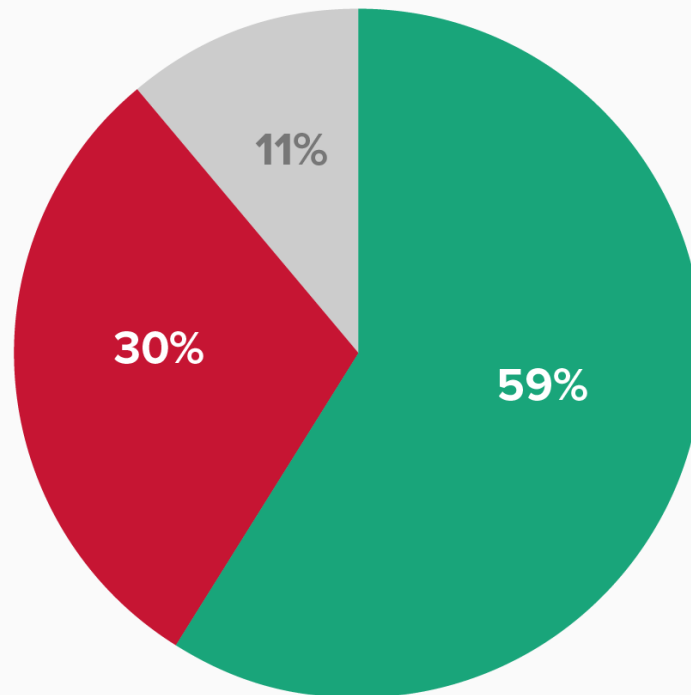




# Most Want Government to Regulate Social Media Content

Do you agree that the government should play a role in regulating the content moderation policies social media companies create for their sites?

● Agree ● Disagree ● Don't know/No opinion



## Muitas nuances, poucos consensos!

Content Regulation	Account Regulation	Visibility Reductions (by acct or item)	Monetary (by acct or item)	Other
<ul style="list-style-type: none"> <li>Remove content</li> <li>Suspend content</li> <li>Relocate content</li> <li>Edit/redact content</li> <li>Interstitial warning</li> <li>Add warning legend</li> <li>Add counterspeech</li> <li>Disable comments</li> </ul>	<ul style="list-style-type: none"> <li>Terminate account</li> <li>Suspend account</li> <li>Suspend posting rights</li> <li>Remove credibility badges</li> <li>Reduced service levels (data, speed, etc.)</li> <li>Shaming</li> </ul>	<ul style="list-style-type: none"> <li>Shadowban</li> <li>Remove from external search index</li> <li>Nofollow authors' links</li> <li>Remove from internal search index</li> <li>Downgrade internal search visibility</li> <li>No auto-suggest</li> <li>No/reduced internal promotion</li> <li>No/reduced navigation links</li> <li>Reduced virality</li> <li>Age-gate</li> <li>Display only to logged-in readers</li> </ul>	<ul style="list-style-type: none"> <li>Forfeit accrued earnings</li> <li>Terminate future earning (by item or account)</li> <li>Suspend future earning (by item or account)</li> <li>Fine author/impose liquidated damages</li> </ul>	<ul style="list-style-type: none"> <li>Educate users</li> <li>Assign strikes/warnings</li> <li>Outing/unmasking</li> <li>Report to law enforcement</li> <li>Put user/content on blocklist</li> <li>Community service</li> <li>"Restorative justice"/apology</li> </ul>

# Moderação de Conteúdo em Escala

Subsídios para análise a partir dos frameworks de Security and Threats Analysis



Government  
Communication  
Service

2019

# RESIST

Counter-disinformation toolkit

Source: <https://gcs.civilservice.gov.uk/publications/resist-counter-disinformation-toolkit/>

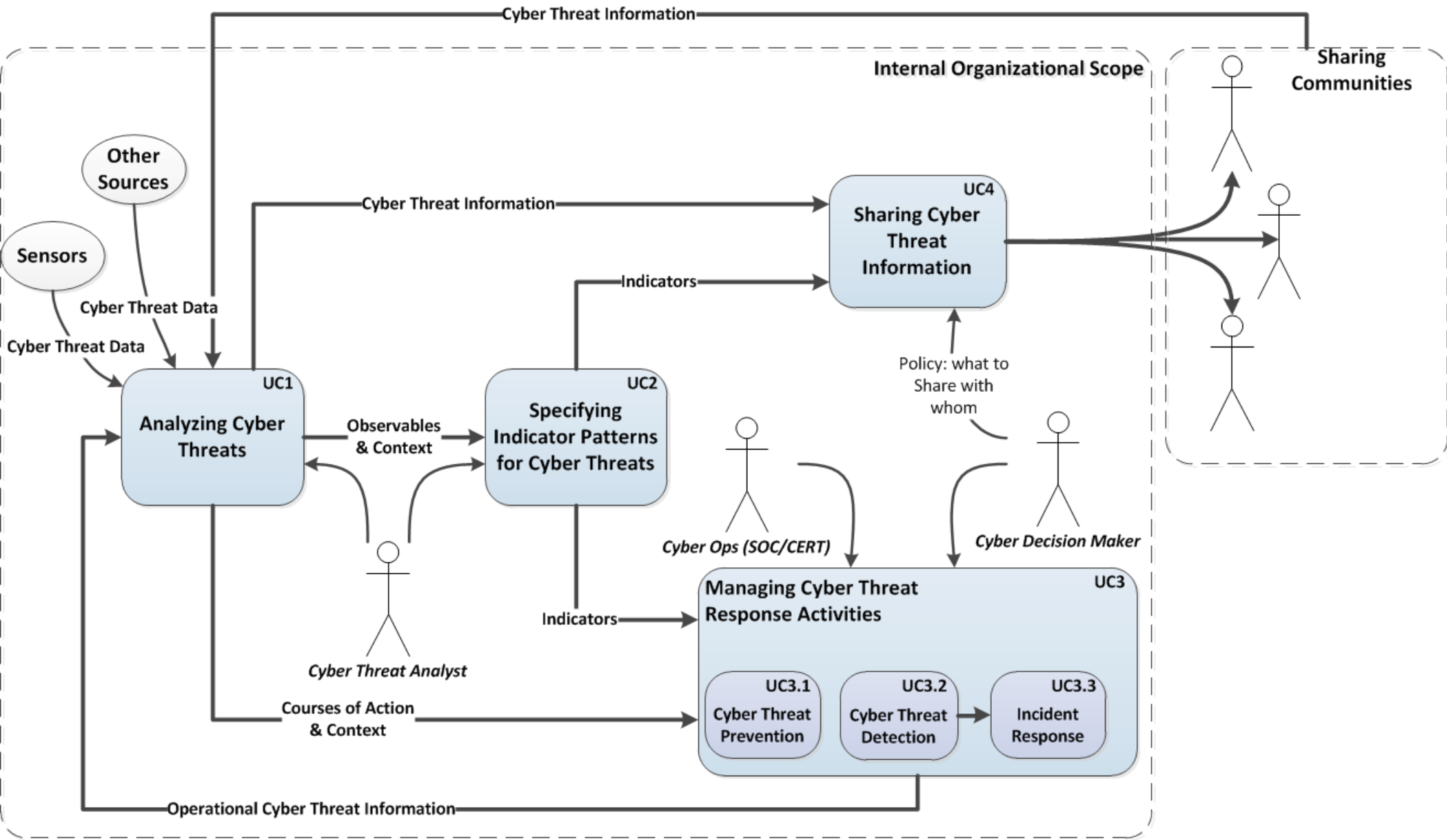
<p><b>DEEPPAKES (F, I, T)</b> Use of digital technology to fabricate facial movements and voice, sometimes in real time.</p>	<p>A fabricated video of a politician shows them saying something outrageous or incriminating, with the goal of undermining confidence in government.</p>
<p><b>ECHO CHAMBER (S)</b> A situation where certain ideas are reinforced by repetition within a social space online.</p>	<p>Creation of internet sub-groups, often along ideological lines, where people engage with like-minded people, which reinforces pre-existing beliefs.</p>
<p><b>FAKE NEWS (F)</b> Deliberate disinformation disguised as news.</p>	<p>A non-journalist fabricates a news story to influence public opinion and to undermine the credibility of mainstream media, which is published on a private platform.</p>
<p><b>FAKE PLATFORM (I)</b> Identity of a web platform is disguised to promote fabricated content.</p>	<p>A web platform is designed to appear like an official site, with the goal of creating the appearance of a credible source of information.</p>
<p><b>FILTER BUBBLE (I, T)</b> Algorithms which personalise and customise a user's experience on social media platforms might entrap the user in a bubble of his or her own making.</p>	<p>The social media flow of a user interested in Brexit gradually adapts to consumed content to eventually only show information in favour of Brexit.</p>
<p><b>FLOODING (T)</b> The overflowing of a target media system with high-volume, multi-channel disinformation.</p>	<p>Multiple commentators, both in the form of bots and real users, make an overwhelming amount of posts with nonsense content to crowd out legitimate information.</p>
<p><b>FORGERY (F, I)</b> Product or content is wholly or partly fabricated to falsely ascribe the identity of the source.</p>	<p>A false document with an official-looking government heading is produced to discredit the government.</p>
<p><b>HACKING</b> Use of illegitimate means to unlawfully gain access to, or otherwise disturb the function of, a platform.</p>	<p>An actor illegitimately claims access to a network from which private information, such as emails, is extracted.</p>
<p><b>HIJACKING (S, T)</b> Unlawful seizure of a computer or an account.</p>	<p>A website, hashtag, meme, event or social movement is taken over by an adversary or someone else for a different purpose.</p>
<p><b>LAUNDERING (F, I)</b> The process of passing of disinformation as legitimate information by gradually distorting it and obscuring its true origin.</p>	<p>A false quote is referenced through multiple fake media channels until the original source is obscured and the quote is accepted as real by legitimate actors.</p>

<p><b>LEAKING (S, T)</b> Disseminating unlawfully obtained information.</p>	<p>Unlawfully obtained emails are leaked to compromise individual actors or to undermine public confidence.</p>
<p><b>MALIGN RHETORIC (R)</b> Lingual ruses aimed at undermining reasonable and legitimate debate and silencing opinions.</p> <ul style="list-style-type: none"> <li>- <b>NAME CALLING (R)</b> A classic propaganda technique based on abusive or insulting language directed against a person or a group.</li> <li>- <b>AD HOMINEM (R)</b> Argumentative strategy focused on attacking the person making the argument rather than the content of the argument itself.</li> <li>- <b>WHATABOUTERY (R)</b> A rhetorical manoeuvre which discredits an opponent's position by accusing them about unrelated issues.</li> <li>- <b>GISH GALLOP (R)</b> A debate tactic focused on drowning the opponent in an overwhelming amount of weak arguments which require great effort to rebut as a whole.</li> <li>- <b>TRANSFER (R)</b> A classic propaganda technique based on transferring blame or responsibility to associate arguments with admired or despised categories of thought.</li> <li>- <b>STRAWMAN (R)</b> A form or argument which targets and refutes an argument that has not been present in the discussion.</li> </ul>	<p>A combination of different rhetorical moves is applied in online conversation to ridicule and diminish other opinions.</p>
<p><b>MANIPULATION (F)</b> Alteration of content to change its meaning.</p>	<p>An image is cropped to only show some of the participating parties in an incident.</p>
<p><b>MISAPPROPRIATION (I)</b> Falsely ascribing an argument or a position to another's name.</p>	<p>A public figure is incorrectly cited or falsely attributed as a source.</p>
<p><b>PHISHING (I, T)</b> A method to unlawfully obtain information online via malware distributed over emails or web platforms.</p>	<p>Malicious links are distributed via email which lead to phishing sites.</p>
<p><b>POINT AND SHRIEK (S)</b> Exploitation of sensitivity to perceived injustices in society to create outrage.</p>	<p>A commentator diverts from a real issue at hand by pointing out the audacity of a make-belief incident which play on pre-existing social grievances.</p>










## Glossary of disinformation techniques

Technique	Example
<p><b>ASTROTURFING (I)</b> Falsely attributing a message or an organisation to an organic grassroots movement to create false credibility.</p>	A source pays or plants information that appear to originate organically or as a grassroots movement.
<p><b>BANDWAGON EFFECT (S)</b> A cognitive effect where beliefs increase in strength because they are shared by others.</p>	A person is more willing to share an article when seeing it is shared by many people.
<p><b>BOT (I, T)</b> Automated computer software that performs repetitive tasks along a set of algorithms.</p> <ul style="list-style-type: none"> <li>- <b>IMPERSONATOR BOTS (I, T)</b> Bots which mimic natural user characteristics to give the impression of a real person.</li> <li>- <b>SPAMMER BOTS (I, R, T)</b> Bots which post repeat content with high frequency to overload the information environment.</li> </ul>	Bots can be used to amplify disinformation or to skew online discussion by producing posts and comments on social media forums and other similar tasks – sometimes they focus on quantity and speed (spammer bots); other times they attempt to mimic organic user behaviour (impersonator bots) – bots can also be used for hacking and to spread malware.
<p><b>BOTNET (I, T)</b> A network of hijacked computers used to execute commands.</p>	Infests personal computers with malware, contribute to DDoS attacks, and distributing phishing attacks.
<p><b>CHEERLEADING (R)</b> The overwhelming promotion of positive messages.</p>	A dissenting opinion is crowded out by positive messages perpetuated by an abundance of commentators cheerleading the ‘right’ opinion.
<p><b>DARK ADS (F, T)</b> Targeted advertisement based on an individual user’s psychographic profile, ‘dark’ insofar as they are only visible to targeted users.</p>	An advertisement containing false information targeted to social media users with personal traits deemed susceptible to this messaging, the goal of shaping their opinions in a specific direction.
<p><b>DDoS ATTACKS (T)</b> Distributed Denial of Service (DDoS) is a cyber-attack where multiple IP addresses are used to disrupt services of a host connected to the internet.</p>	A DDoS attack is conducted to bring down a government website during a crisis, to deny citizens access to reliable information.

<p><b>POTEMKIN VILLAGE (I, R)</b> A smoke-screen of institutions and/or platforms established to deceive audiences.</p>	A complex network of fake think tanks is established to disseminate disinformation which seems legitimate due to the perceived legitimacy of the network.
<p><b>RAIDING (S, T)</b> Temporarily disrupting a platform, event, or conversation by a sudden show of force.</p>	Several automated accounts are coordinated to disrupt a conversation by temporarily spamming nonsense messages.
<p><b>SATIRE AND PARODY (R, S)</b> Ridiculing and humouring of individuals, narratives or opinions to undermine their legitimacy.</p>	A public figure is ridiculed using memes where non-factual opinions are ascribed to the public figure.
<p><b>SHILLING (I)</b> To give credibility to a person or a message without disclosing intentions or relationships.</p>	An actor endorses certain content while appearing to be neutral but is in fact a dedicated propagandist.
<p><b>SOCKPUPPETS (I, R, T)</b> Use of digital technology to disguise identity, to play both sides of a debate.</p>	A user creates two or more social media accounts under opposing identities i.e. one pro-fox hunting, one against, with the aim of playing the identities against one another.
<p><b>SPIRAL OF SILENCE (S)</b> The decrease in audibility of deviant opinions due to non-conforming beliefs.</p>	A person with non-conforming minority beliefs is less willing to share his or her opinions.
<p><b>SYMBOLIC ACTION (S)</b> Refer to acts that carry symbolic value in the sense that they signal something to an audience to create a response.</p>	A user plays on universally shared symbolic cues e.g. terrorist attacks to create a climate of fear.
<p><b>TAINTING (F, S, T)</b> Leaked contents are tainted with forgeries.</p>	Leaked documents are distributed together with carefully placed fakes.
<p><b>TERRORISM (R, S)</b> Imagery from real-world events is used to make political claims.</p>	Images of violence are used to support false narratives, with the aim of creating a climate of fear or justifying a political argument.
<p><b>TROLLING (I, R, S)</b> Deliberate commenting on internet forums to provoke and engage other users in argument.</p>	Social media users deliberately post provocative comments to create emotional outrage in other users.
<p><b>WOZZLE EFFECT (R)</b> Self-perpetuating evidence by citation.</p>	A false source is cited repeatedly to the point where it is believed to be true because of its repeated citation.



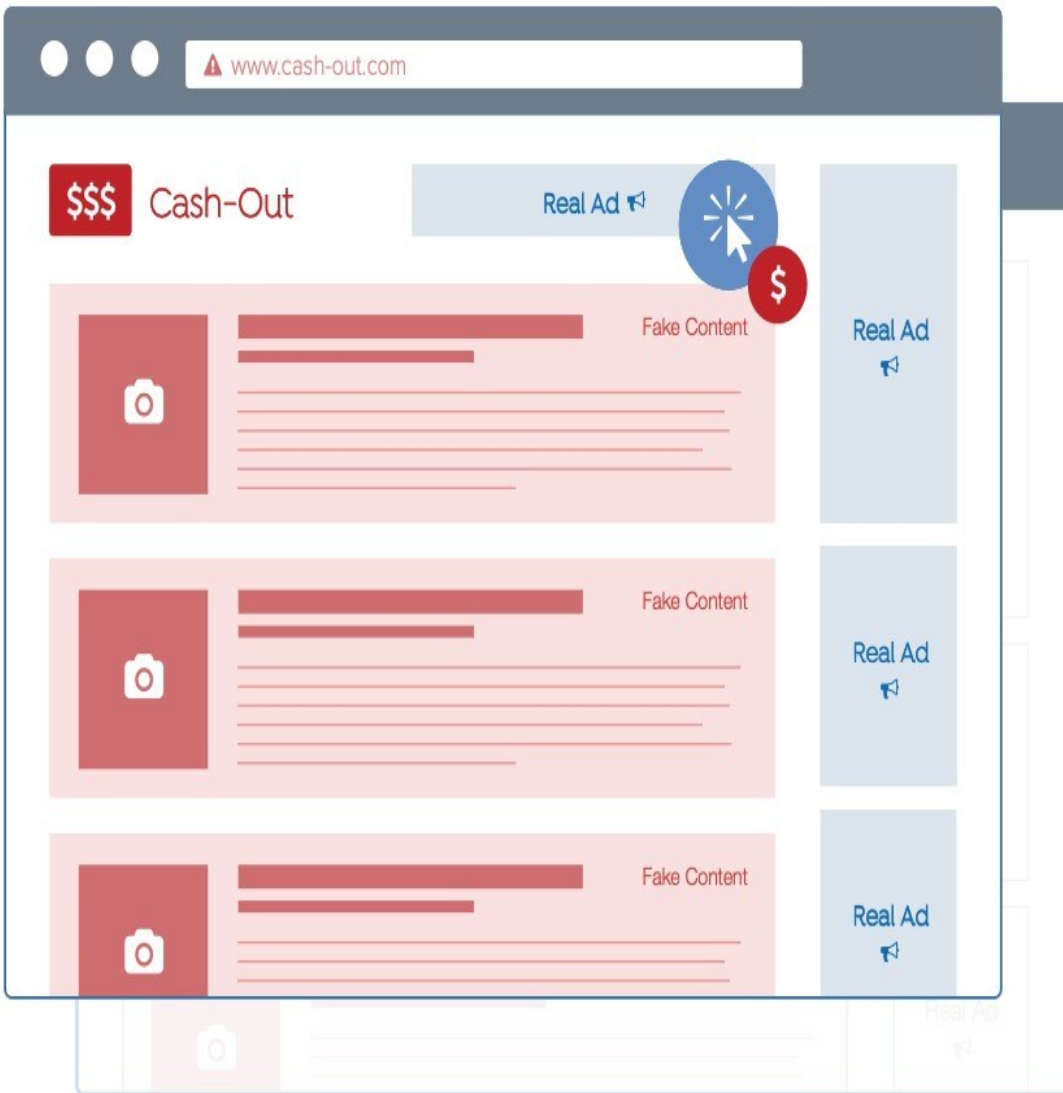
STIX 2.1 defines 18 STIX Domain Objects (SDOs):

Object	Name	Description
	<b>Attack Pattern</b>	A type of TTP that describe ways that adversaries attempt to compromise targets.
	<b>Campaign</b>	A grouping of adversarial behaviors that describes a set of malicious activities or attacks (sometimes called waves) that occur over a period of time against a specific set of targets.
	<b>Course of Action</b>	A recommendation from a producer of intelligence to a consumer on the actions that they might take in response to that intelligence.
	<b>Grouping</b>	Explicitly asserts that the referenced STIX Objects have a shared context, unlike a STIX Bundle (which explicitly conveys no context).
	<b>Identity</b>	Actual individuals, organizations, or groups (e.g., ACME, Inc.) as well as classes of individuals, organizations, systems or groups (e.g., the finance sector).
	<b>Indicator</b>	Contains a pattern that can be used to detect suspicious or malicious cyber activity.
	<b>Infrastructure</b>	Represents a type of TTP and describes any systems, software services and any associated physical or virtual resources intended to support some purpose (e.g., C2 servers used as part of an attack, device or server that are part of defence, database servers targeted by an attack, etc.).
	<b>Intrusion Set</b>	A grouped set of adversarial behaviors and resources with common properties that is believed to be orchestrated by a single organization.
	<b>Location</b>	Represents a geographic location.

Source: <https://oasis-open.github.io/cti-documentation/stix/intro>



# Best Practices



<https://www.whiteops.com/blog/the-shoe-is-a-lie-how-an-android-botnet-defrauded-advertisers-and-consumers>

## The Hunt for 3ve

Taking down a major ad fraud operation through industry collaboration

November 2018

Source:

<https://resources.whiteops.com/research-and-investigations/the-hunt-for-3ve>

## CASH-OUT SITE

Un sitio web, aplicación u otro recurso que es capaz de entregar anuncios, y es operado por ciberdelincuentes con el propósito de extraer dinero del ecosistema de publicidad en internet.

Co-authored by Google and White Ops  
with technical contributions by Proofpoint and others



# SIM Farms: The Modern Trojan for Mobile Operators

TRENDING TOPICS APRIL 12, 2017

Source: <https://haud.com/blog/2017/04/12/sim-farms-modern-trojan-mobile-operators/>

## Definitions

### *Spamming*

Unwanted messages delivered to subscribers

### *Flooding*

Massive amount of messages sent to nodes and subscribers

### *Faking*

The illegal use of SMSC identity by a foreign system

### *Spoofing*

Messages sent illegally by simulating a roaming subscriber

### *Smishing*


Deceptive messages attempting to acquire subscriber information

### *Virus distribution*

Messages luring subscribers to a download site with viruses

# Calls + SMS, +55 Brazil

With the Call + SMS service you make the best decision. Receive and make calls, send and receive SMS with one single virtual number. Get a virtual SIM card for calls and text messages. 🤖 Free phone number - [Register now](#) and get your first virtual phone number for free.

 Brazil \* ▾  ▾  ▾

## Phone numbers in Brazil

[Virtual numbers](#) / Brazil

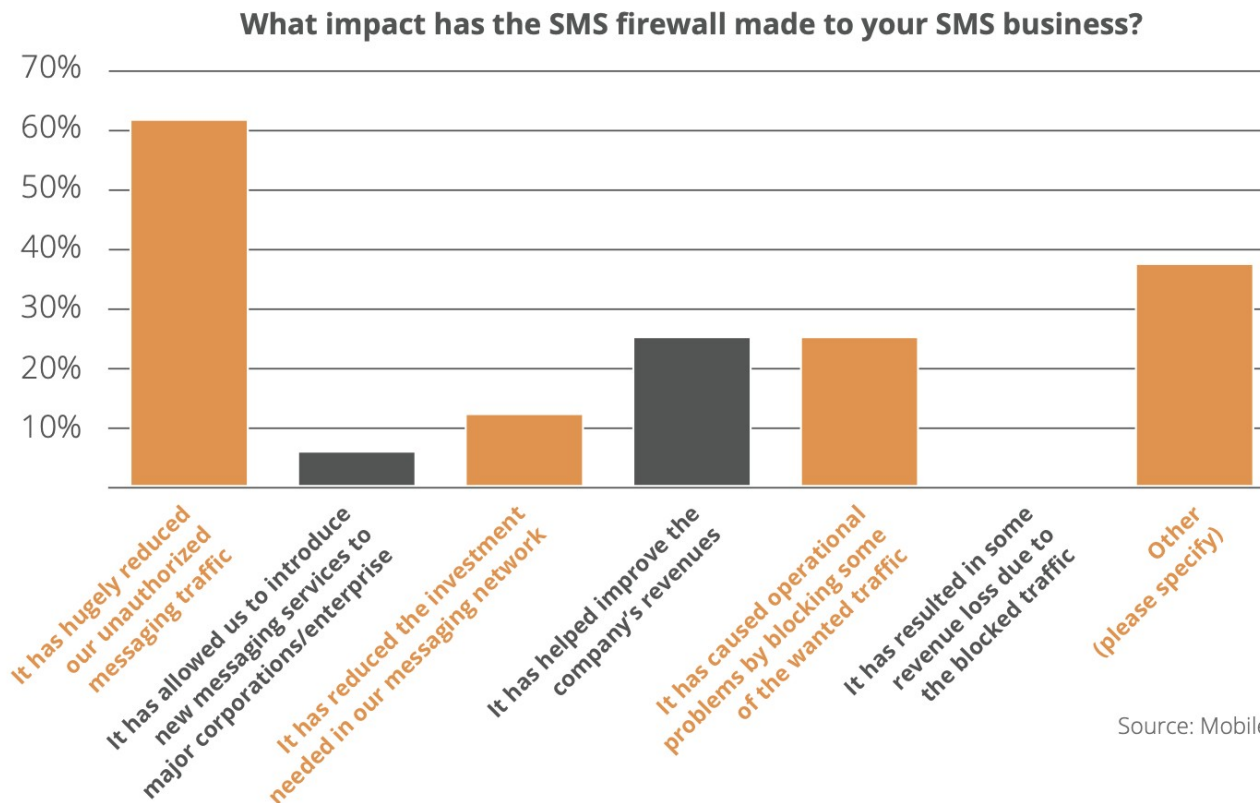
Zadarma gives you the opportunity to connect a phone number in Brazil to your PC, SIP gate, office PBX, mobile phone, or to any other device that supports SIP. In addition, you can forward your virtual phone number to any country for free, or at a very **low price**.

You can get phone numbers in the cities listed below at the following prices:

Area code	Destination	Connection fee	Monthly fee	
021	<b>Mobile</b> <sup>1</sup>	\$0	\$8	<b>BUY</b>
0800	<b>Toll-free</b> <sup>2</sup>	\$0	\$6	<b>BUY</b>
081	<u>Abreu E Lima</u> <sup>3</sup>	\$0	\$5	<b>BUY</b>
041	<u>Almirante Tamandare</u> <sup>3</sup>	\$0	\$5	<b>BUY</b>
011	<u>Alphaville</u> <sup>3</sup>	\$0	\$5	<b>BUY</b>
051	<u>Alvorada</u> <sup>3</sup>	\$0	\$5	<b>BUY</b>
019	<u>Americana</u> <sup>3</sup>	\$0	\$5	<b>BUY</b>

*<25% of MNOs have invested in the necessary SMS firewall required to transform grey routes into white routes*

# Firewall Impact

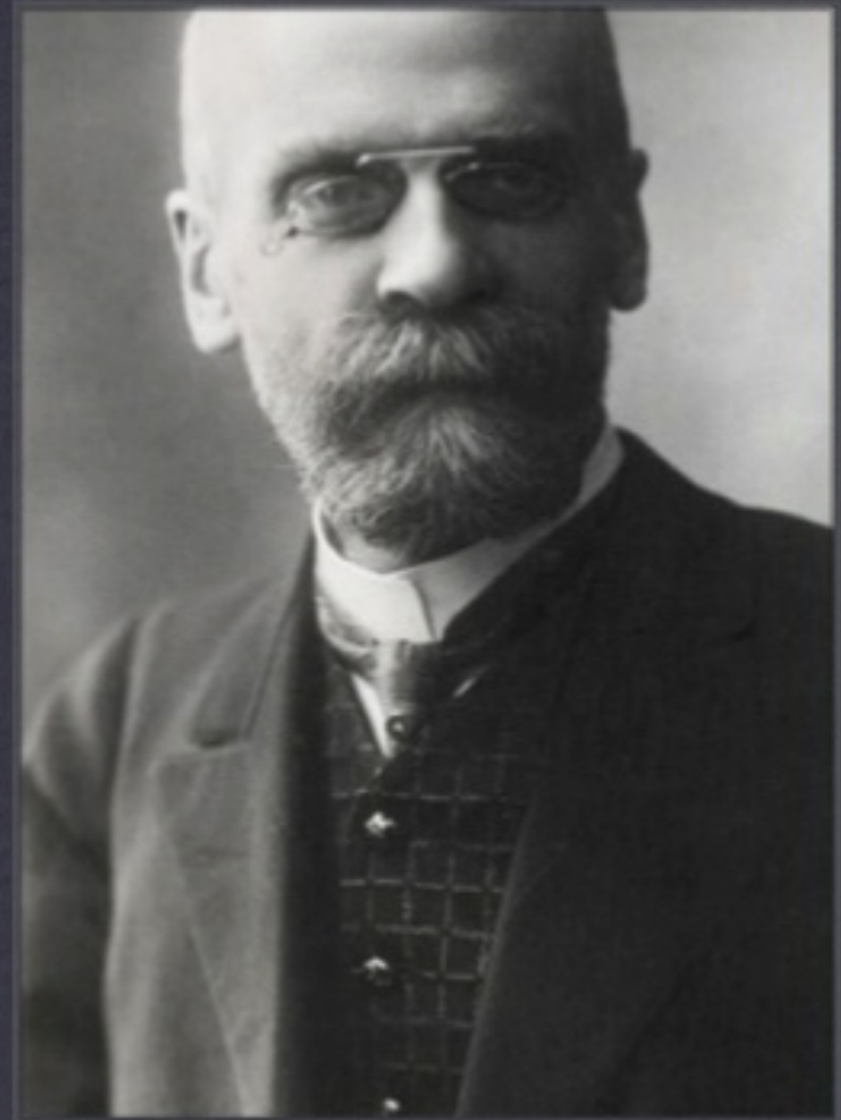


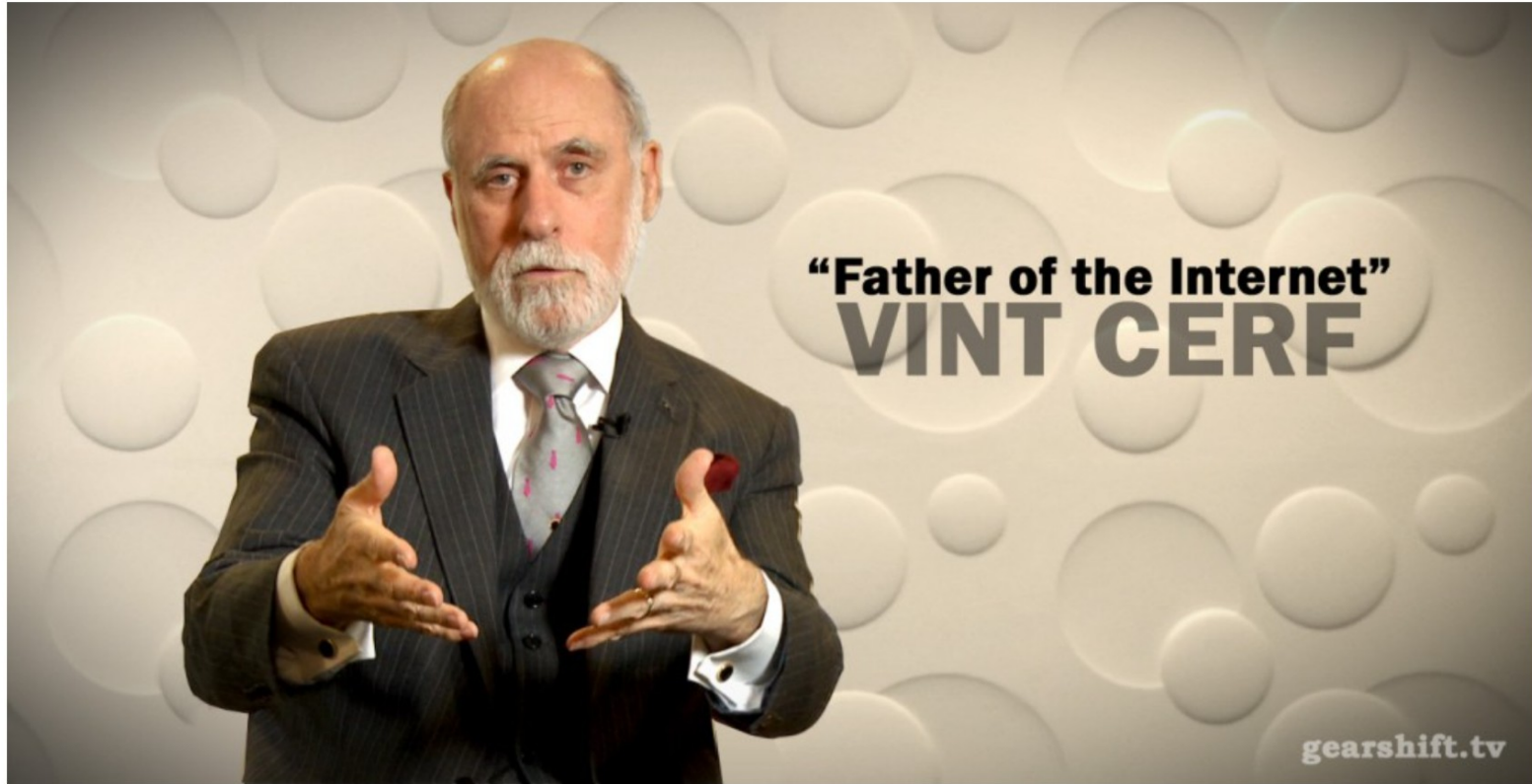
Source: MobileSquared

À Guisa de Conclusão

Onde houver sociedade,  
haverá crime

**EMILE  
DURKHEIM**  
**1858-1917**





**“The internet is a reflection of our society and that mirror is going to be reflecting what we see. If we do not like what we see in that mirror the problem is not to fix the mirror, we have to fix society.”**

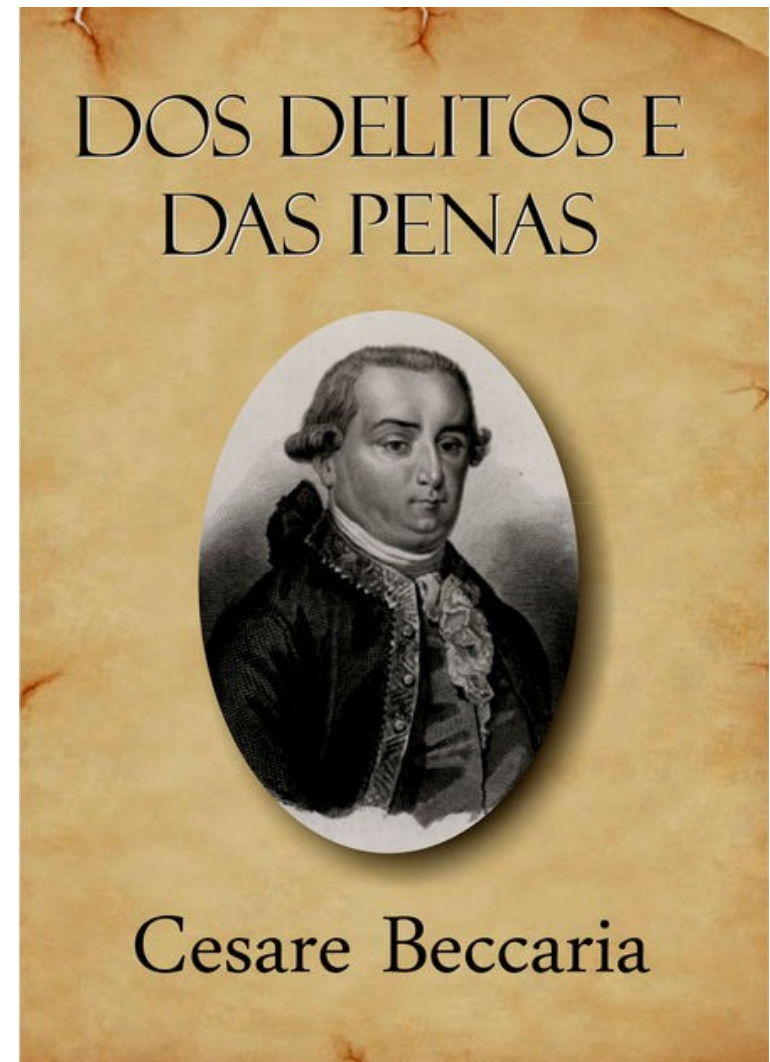
Vint Cerf

**“A internet é um reflexo da nossa sociedade e esse espelho vai refletir o que vemos. Se não gostamos do que vemos nesse espelho, o problema não é consertar o espelho, temos de consertar a sociedade.”**

Vint Cerf

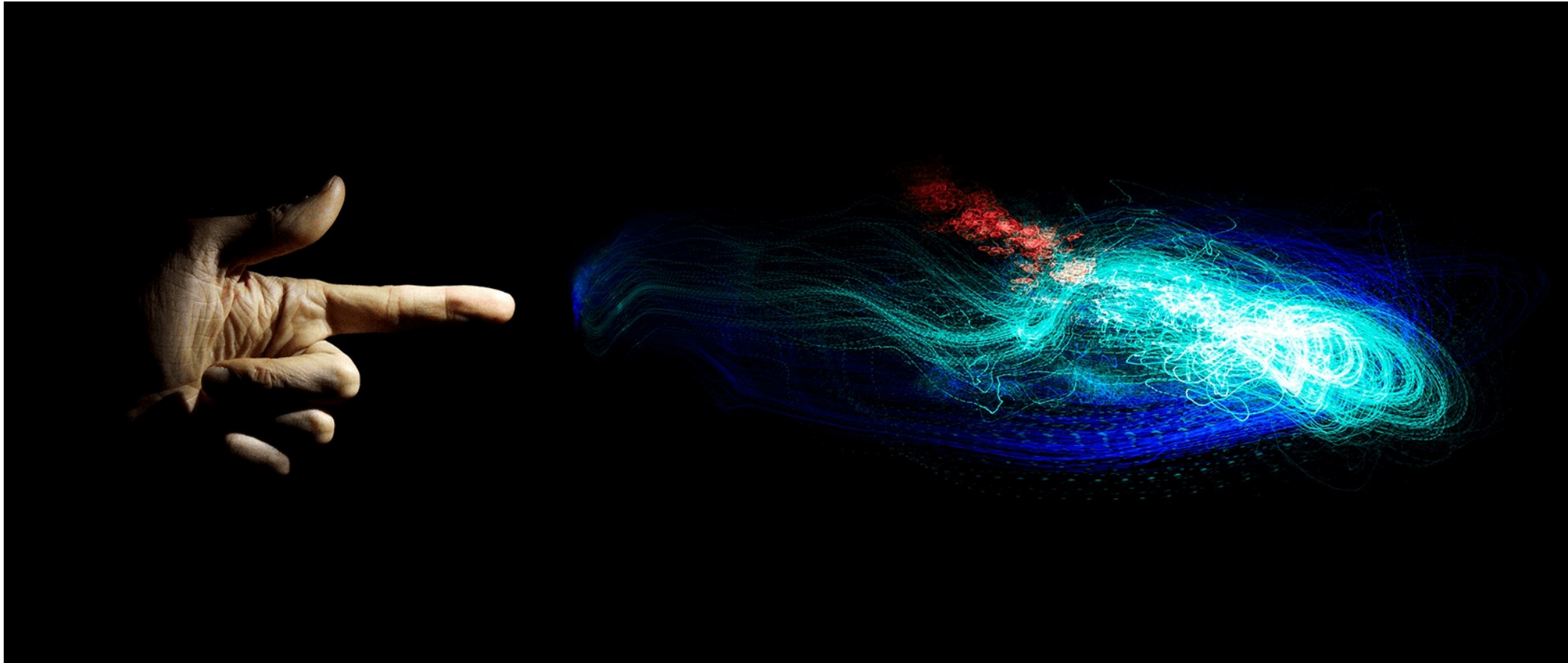
**"É melhor prevenir os crimes do que ter de puní-los. O meio mais seguro, mas ao mesmo tempo mais difícil, de tornar os homens menos inclinados a praticar o mal é aperfeiçoar a educação"**

***In: BECCARIA, Cesare Bonesana. Dei delitti e delle pene: Milão, 1764.***





# Não existe bala de prata!



Foco em estratégias multisetoriais:

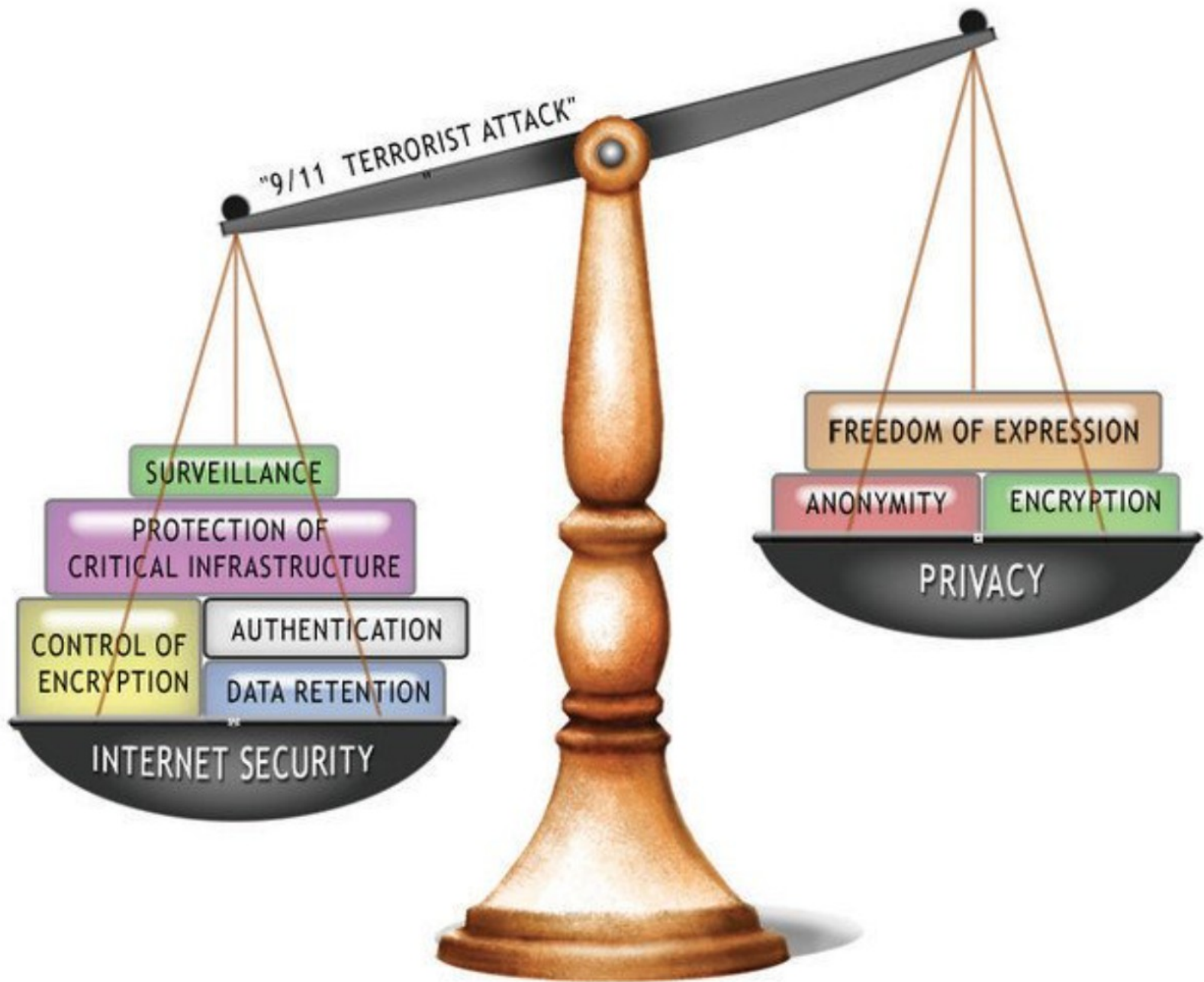
- a) detecção, resposta, transparência e accountability (curto prazo)
- b) estratégias de dissuasão (prevenção geral) + frustrar o resultado (post facto)
- c) educação para o uso ético, seguro e responsável da Internet (longo prazo)

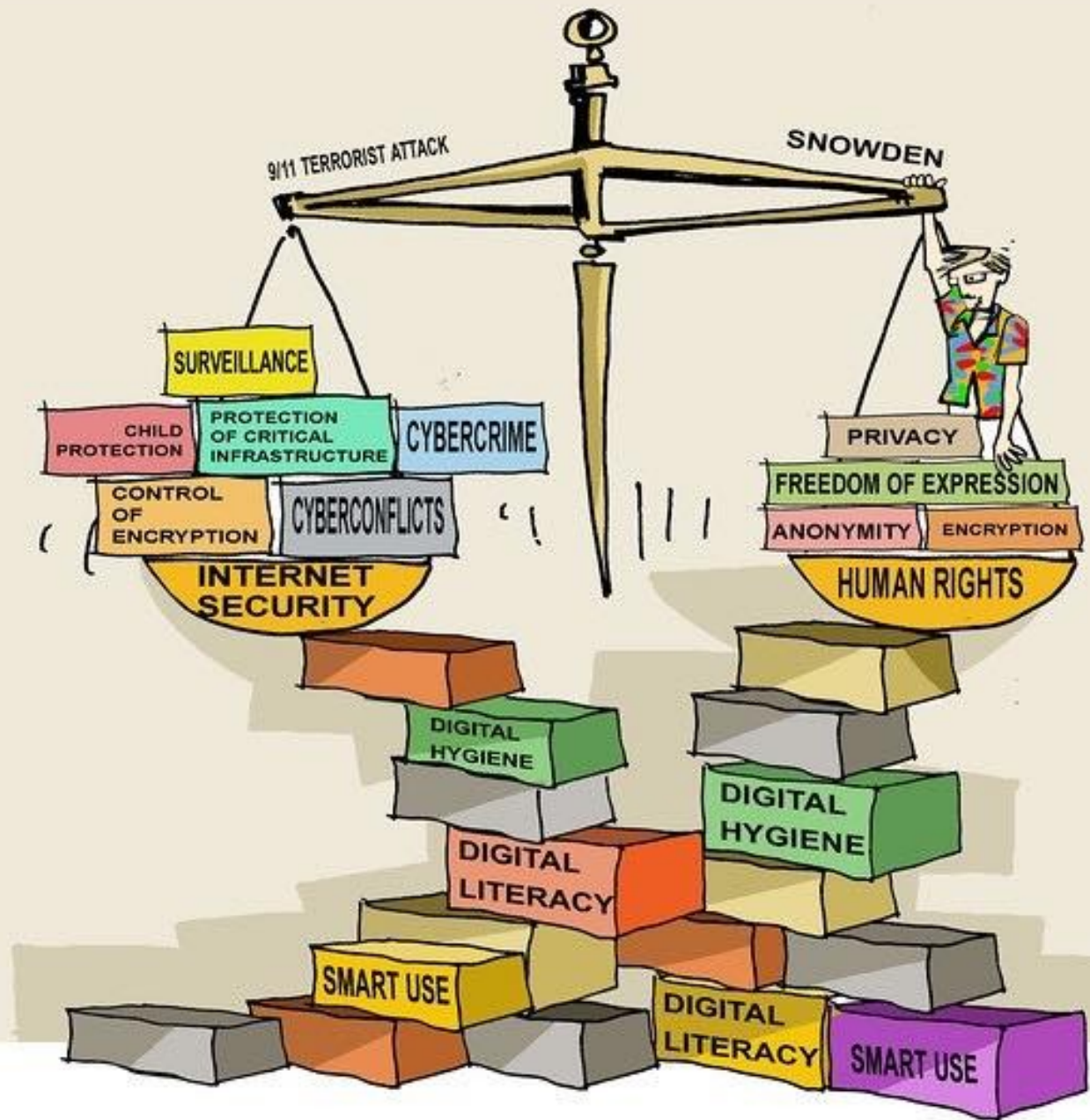


## **The policy response to fake news: China as a comparative case**

The example of China's crackdown on "online rumours" since 2013 is a useful illustration of the dangers of (i) establishing structures of prepublication regulation and (ii) having too wide definition of what constitutes unverified fake news or "rumour".

During 2013-2014 it was reported that the Chinese authorities had intensified their policy of deleting posts on Chinese social media such as Wechat. Chinese authorities claimed that these were "necessary to safeguard citizens' rights and interests, and promote the healthy development of the internet". The Chinese approach is to make operators of social networks responsible for removing a widely defined category of content considered to be 'rumours' and jail terms of up to 3 years for those responsible. Service providers are required to suspend the accounts of those found to be responsible for spreading "irresponsible rumours". A number of categories of such rumours are identified: these include undermining morality, the socialist system, and the authenticity of information. Discretion for deciding what fits into these categories lies with the social networks, but these are periodically reviewed under the terms of their licences. In China, social networks must be in receipt of several different licences from central government.<sup>h</sup> Anecdotal evidence suggests that incentivising intermediary filtering and blocking through the threat of strong penalties leads to intermediaries developing automated blocking and filtering, together with expensive human-led programs of deletion. Due to the lack of transparency it is impossible to know precisely what is blocked, but the evidence reported by Western journalists suggests that over blocking is rife.





Concept: Vladimir Radunović Illustration: Vladimir Veljašević

**OBRIGADO!**



[thiagotavares@safernet.org.br](mailto:thiagotavares@safernet.org.br)