Menaces informationnelles : le développement du « Trust & Safety » dans les organisations

Caroline RABOURDIN
Doctorante sur les guerres de l'information en Sciences de l'information et de la communication
Programme doctoral Défense-Sécurité intérieure
Université Aix-Marseille

Alors que les départements Trust & Safety sont connus aux Etats-Unis, a contrario en France, la majorité des entreprises ignorent tout de leur existence.

Né aux Etats-Unis, au début des années 2000 sous l'impulsion des géants du numérique comme eBay, Google et Facebook pour garantir la confiance des utilisateurs et la sécurité des réseaux, le « Trust & Safety » (« Confiance & Sécurité ») désigne aujourd'hui un ensemble de pratiques regroupées au sein d'un domaine d'activité devenu un département d'entreprise au même titre que le marketing, les ressources humaines ou la comptabilité.

Si les plateformes numériques ont été les premières à mettre en place des départements Trust & Safety, d'autres secteurs ont depuis emboîté le pas, comme les fintechs, marketplaces, jeux en ligne, assurances, services de santé, mobilité, institutions, etc.

Face à la multiplication des cyberattaques, à l'utilisation malveillante croissante des intelligences artificielles et à l'essor des campagnes de désinformation, les entreprises sont confrontées à des risques multiples et protéiformes. Elles doivent sans cesse s'adapter pour répondre aux attaques et garantir la sécurité des personnes, systèmes et données, tout en préservant leur réputation et en respectant les lois, les règlementations et la vie privée de chacun. Le Trust & Safety apparaît donc comme une parade pour contrer les menaces informationnelles diverses.

Un domaine transversal et pluridisciplinaire

Le Trust & Safety est à la fois un ensemble de procédés, une fonction transversale et un domaine d'activité. En entreprise, la taille des équipes Trust & Safety varie considérablement d'une organisation à une autre. En mars 2024, Google comptait environ 250 collaborateurs, tandis que chez X (ex-Twitter), l'équipe Trust & Safety était composée de 2 300 personnes. Quant à Meta et TikTok, ils ont chacun annoncé employer près de 40 000 personnes.

Le Trust & Safety consiste à prévenir, détecter, protéger, répondre aux manipulations de l'information comprenant entre autres cyberattaques, désinformation, contenus trompeurs ou dangereux, fraudes, usurpations d'identité, etc. Il consiste à prévenir les atteintes psychologiques (harcèlement, menaces…), physiques (appels à la violence…), légales (contenus extrémistes, pédocriminels…) et économiques (escroqueries, ransomwares…).

Pluridisciplinaire, le Trust & Safety regroupe gestion

des risques, cybersécurité, compliance (ou conformité), gestion des contenus, intelligence économique, communication, design produit, droit, expérience utilisateur (UX). Il rassemble ainsi experts cyber, modérateurs, analystes, juristes et avocats, data scientists, ingénieurs, responsables produits, rédacteurs, détectives...

Les équipes analysent les comportements des utilisateurs, détectent des schémas suspects (« pattern recognition »), répondent aux incidents, alertent ou sensibilisent les clients, élaborent des politiques internes, rédigent des rapports de transparence, renforcent leurs protocoles, collaborent avec différentes parties prenantes, etc. Leur approche est aussi proactive. En effet, le T&S ne se contente pas de réagir aux attaques : il anticipe également les menaces émergentes grâce à une veille stratégique et technologique, et finance des recherches scientifiques sur le sujet.

Quelles sont les missions du Trust & Safety ?

Gestion des risques

La démarche Trust & Safety commence par une analyse des risques spécifique. Il ne s'agit pas là d'une gestion des risques complète à l'organisation mais bien dédiée aux manipulations de l'information. Cette étape vise à identifier les vulnérabilités liées à la sécurité des données, des personnes, des infrastructures numériques et des flux informationnels internes comme externes. Puis, une stratégie de gestion des risques identifiés est déployée. Elle comprend notamment la mise en place de mesures préventives (protocoles de sécurité, formation, veille…) et curatives (réponse aux incidents…).

Cybersécurité

Le Trust and Safety inclut également la cybersécurité, permettant de protéger l'entreprise et ses utilisateurs contre les brèches de sécurité et accès non autorisés, les malwares (virus, spywares, ransomwares…), les tentatives de phishing, le vol, l'exfiltration ou la corruption de données, les utilisations malveillantes des intelligences artificielles (deepfakes, automatisation d'attaques, etc.).

Cela passe par des audits de sécurité et divers « pentest » (tests de pénétration) et le renforcement des systèmes d'information et des dispositifs de détection des anomalies. Les équipes doivent ensuite s'assurer que les protocoles et les solutions déployées sont bien compliants avec les réglementations en vigueur (RGPD, DSA, NIS2, etc.). Enfin, les experts cyber sensibilisent les équipes, clients et parties prenantes sur la sécurité des mots de passe, l'authentification à double facteur et l'hygiène numérique.

Modération des contenus

Aujourd'hui, la majorité des entreprises ont un site Internet, intégrant éventuellement une boutique en ligne, et sont sur les réseaux sociaux. Autant de points d'entrée potentiels pour des acteurs malveillants. Si le Trust & Safety intervient sur la partie cyber, il ne se limite pas pour autant aux infrastructures techniques. Il vise aussi à lutter contre les menaces informationnelles relatives au contenu, dirigées contre l'entreprise, ses salariés, clients, parties prenantes, ainsi que contre ses produits.

Désinformation, propagande, chantage, menaces, discours haineux, apologie du terrorisme, incitation à l'autodestruction, contenu violent, challenges dangereux sont passés au crible, tout comme les avis clients, les commentaires et autres contenus créés par autrui (« user-generated content »). Le but : éviter, limiter et contenir les crises réputationnelles ou sécuritaires. Pas question en effet de laisser passer sur les réseaux sociaux de fausses

informations sur le dirigeant ou de révéler publiquement son adresse, ou encore d'accepter qu'un compte social utilise l'image de la marque, propage des fake news sur l'un de ses produits, ou menace certains salariés par messages privés.

Pour modérer l'ensemble des contenus, les équipes s'appuient sur des outils automatisés et recourent à des interventions humaines pour évaluer le second degré, l'humour et les références culturelles. Elles peuvent également être amenées à travailler en partenariat avec les gouvernements dans le cadre de leur lutte contre la désinformation, lors de guerres hybrides notamment. Pour les entreprises, la modération peut aussi inclure un volet Diplomatie digitale afin d'échanger avec les plateformes, connaître leurs règles de modération ou demander la suppression de certains contenus malveillants.

Management de l'escalade

En cas de plaintes de clients sur les réseaux sociaux, de conflit avec l'organisation ou d'amplification d'une crise réputationnelle, le département gère l'escalade en mettant en place des stratégies de réponses pour réagir de manière proportionnée, rapide, efficace, tout en respectant les valeurs de l'entreprise, ses conditions générales (« policies ») ainsi que le cadre légal. Les équipes implémentent des systèmes de détection et des protocoles de signalement, attribuent des pénalités (« red flag »), suppriment tout contenu portant atteinte, bloquent les comptes et recourent à des actions juridiques, si besoin.

L'objectif est double : garantir un espace numérique sûr pour tous et maintenir un climat de confiance avec la marque.

Intelligence économique

Lors d'une attaque informationnelle, le département T&S mobilise ses analystes pour mener des investigations et remonter à son origine. Il intervient également lors de fraudes, d'escroqueries, d'usurpations d'identité, de vol. Les experts analysent les comportements suspects et schémas d'activité, identifient les acteurs malveillants, cartographient les réseaux d'attaquants et attribuent des responsabilités. En collaboration avec des détectives, ils collectent les éléments qui serviront plus tard de preuves devant les tribunaux.

Selon le domaine d'activité de l'entreprise, les équipes Trust & Safety peuvent être amenées à identifier les campagnes d'influence étrangères, assurer le suivi de certains posts, suivre une tendance, afin d'aider les gouvernements dans leur lutte contre les ingérences.

Enfin, les équipes Trust & Safety mènent une veille stratégique et technologique pour anticiper les menaces et concevoir des outils de détection et de protection adaptés.

Politique d'utilisation - « Policy »

Le Trust & Safey inclut la rédaction et la mise à jour de conditions générales ou politiques d'utilisation. Ces codes de conduite définissent les règles de ce qui est acceptable et ce qui ne l'est pas. Ils concernent la modération des contenus, la gestion des données personnelles, la résolution de conflits, etc. Ces conditions peuvent prendre la forme de chartes, standards, procédures, règlements. Elles évoluent en permanence afin de s'adapter aux nouvelles menaces et tactiques des attaquants, et servent de guide de conduite aux modérateurs, notamment.

Application de la loi - « Law enforcement »

En cas d'activités malveillantes, juristes et avocats interviennent pour faire appliquer les conditions générales d'utilisation et les obligations légales (RGPD, DSA, Section 230…). Ils participent aussi à la

définition de protocoles de réponse aux violations des CGU, aux fraudes, aux actes de contrefaçons ainsi qu'aux incidents de cybersécurité. Ils défendent également la propriété intellectuelle et le droit à l'image, et luttent par exemple contre l'utilisation non autorisée des licences, du nom de la marque ou de l'image de ses produits. En cas de besoin, une collaboration étroite avec les autorités est engagée lors d'enquêtes.

La transparence

L'obligation de communiquer de manière transparente incombe également à ce département. Dans ses rapports, qu'ils soient internes ou publics, l'organisation doit expliciter ses règles, ses politiques de modération, ainsi que ses conditions générales d'utilisation. Ils expliquent comment les données sont collectées, utilisées, protégées, stockées, comment le contenu est modéré et les comportements d'utilisateurs gérés. Selon le domaine d'activité, ces documents détaillent le volume de modération et les demandes des autorités. La société peut aussi rendre publique certaines décisions, en particulier celles liées aux suspensions, aux retraits de contenus ou aux sanctions. En cas de défaillance ou d'incident majeur, elle peut être amenée à publier des excuses et à annoncer des mesures correctives afin de restaurer la confiance.

Vers un « Safety by design » ?

Et si la sécurité était pensée et réfléchie dès le début des projets ? Le « Safety by design » consiste à intégrer l'ensemble des activités du Trust & Safety dès la création d'une entreprise ou de la conception d'un nouveau produit (« Product management ») ou d'un service. Les menaces informationnelles sont ainsi prises en compte dès le départ. Dans son rapport intitulé « Technology and Trust and Safety » publié en 2024, le gouvernement britannique estime que 65 % des organisation incluent les équipes Trust et Safety dès la phase d'idéation.
Cela inclut par exemple le développement de fonctionnalités, d'interfaces et de systèmes visant à protéger à la fois les infrastructures (sites web, plateformes, réseaux, bases de données…) mais aussi l'ensemble des parties prenantes (utilisateurs, clients, salariés, partenaires, fournisseurs…) contre les attaques cyber et celles liées au contenu.
Dans le cadre du « Product Management », le T&S est placé au cœur de chaque étape du cycle de vie d'un produit : de sa création à son utilisation sur le long terme, voire à ses mises à jour successives dans le cas d'un produit numérique.

Mais la démarche dépasse largement le champ du numérique. Elle concerne tout type d'organisation qui cherche à instaurer une relation de confiance durable avec ses usagers.

Enfin, le Trust & Safety inclut aussi l'expérience utilisateur (« User Experience » ou « UX »). Ici, la sécurité va de pair avec l'ergonomie d'un site Internet, d'une application, ou d'un service. Il s'agit de combiner sécurité et expérience client.

Une certification ISO

Les pratiques T&S font à présent l'objet d'une norme ISO. En juin 2025, l'ISO a publié une nouvelle norme internationale intitulée ISO/IEC 25389:2025, aussi appelée « The Safe Framework ». Elle fournit un cadre de recommandations aux organisations proposant un produit ou service numérique à destination du grand public, afin d'identifier, prévenir et maîtriser les risques liés aux contenus et aux comportements. Cette norme propose également des critères d'évaluation pour mesurer l'efficacité des pratiques mises en place.

Quel retour sur investissement ?

Dans son rapport « Technology and Trust and Safety » publié en 2024, le gouvernement

britannique rapporte que même si le T&S permet de limiter les risques au sein d'une organisation, il reste difficile de quantifier le retour sur investissement.

Contrairement à d'autres fonctions en entreprise, il n'existe pas de formule unique : chaque organisation doit définir ses propres indicateurs en fonction de son secteur et de ses risques. Une entreprise pourra par exemple mesurer la baisse de commentaires négatifs, alors qu'une plateforme e-commerce suivra la diminution du taux de fraude. Le ROI mêle résultats tangibles (réduction des fraudes, baisse des coûts juridiques) et résultats intangibles (confiance des utilisateurs, réputation de la marque).

Toutefois, le T&S est bien synonyme de résultats. Ainsi, le gouvernement britannique précise que les entreprises mettant l'accent sur la sécurité des données, la propriété intellectuelle et la confidentialité obtiennent de meilleurs résultats commerciaux et un meilleur retour sur investissement (ROI). De plus, selon l'enquête Trust and Safety Survey 2025 menée par PwC, la mise en place d'une démarche Trust & Safety a un impact mesurable et favorable sur les consommateurs.

Le T&S ne se contente pas d'atténuer les risques informationnels, il augmente aussi les gains et accroît la valeur et le chiffre d'affaires d'une entreprise. Il augmente la confiance, la sécurité et la réputation, trois actifs immatériels qui conditionnent, à long terme, la performance et la croissance d'une entreprise.

Quelques cas concrets

Après un incident majeur en 2011 ayant ébranlé la crédibilité de la plateforme, Airbnb a massivement investi dans des dispositifs de Trust & Safety. L'entreprise a mis en place des systèmes de vérification d'identité, de paiement sécurisé, des services de photographie professionnelle pour valoriser les annonces, ainsi qu'un mécanisme d'avis et un programme de garantie pour les hôtes. Ces innovations ont permis de réduire les risques perçus par les utilisateurs et de créer une infrastructure forte. Après avoir amélioré ses opérations de Trust & Safety, une plateforme d'e-commerce a constaté un taux de 99 % des problèmes résolus en moins de 24 heures.

Grâce à une refonte des processus de support et de modération, une plateforme nord-américaine de covoiturage a permis de réduire de 40 % les demandes répétées des usagers, d'atteindre 93 % de précision dans la résolution des cas et de garantir des délais de traitement inférieurs à 24 heures.

Conclusion

Alors que la mise en place de cellules d'intelligence économique peine encore à voir le jour au sein des entreprises françaises, les départements Trust & Safety pourraient leur permettre de rattraper ce retard.

Le Trust & Safety est une démarche globale qui vise à protéger l'entreprise ainsi que son écosystème, comprenant son patrimoine matériel et immatériel (ressources humaines, actifs financiers, propriété intellectuelle, réputation…). Face à des menaces informationnelles protéiformes et en constante évolution, et dans un contexte où la confiance est un atout clé, le Trust & Safety s'impose comme un levier stratégique pour assurer la pérennité, la sécurité et la compétitivité des organisations.

Merci à Thibault RENARD, expert IE, senior advisor du CyberCercle, qui a supervisé la rédaction de ce texte, ainsi que Fabienne CROP, spécialiste en IE, et Fabrice FROSSARD, Fondateur Faber Content, qui en ont assuré la relecture.

Sources bibliographiques

DeNARDIS, L., & HACKL, A. M. (2015). Internet governance by social media platforms. Telecommunications Policy, 39(9), 761–770. https://doi.org/10.1016/j.telpol.2015.04.003

GILLESPIE, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

GILLESPIE, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2). https://doi.org/10.1177/2053951720943234

GORWA, R., BINNS, R., & KATZENBACH, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). https://doi.org/10.1177/2053951719897945

JHAVER, S., GHOSHAL, S., BRUCKMAN, A., & GILBERT, E. (2018). Online harassment and content moderation: The case of blocklists. ACM Transactions on Computer-Human Interaction, 25(2),12:1–12:33.https://doi.org/10.1145/3185593

KLONICK, K. (2018). The new governors: The people, rules, and processes governing online speech. Harvard Law Review, 131(6), 1598–1670. https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

MATAMOROS-FERNÁNDEZ, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. Information, Communication & Society, 20(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

MOLLICK, E., & TAN, C. (2022). Trust and safety in the age of AI. MIT Sloan Management Review. https://sloanreview.mit.edu/article/trust-and-safety-in-the-age-of-ai/

RIEDL, M. J., & YOUNG, R. M. (2005). An objective character believability evaluation procedure for multi-agent story generation systems. International Journal of Human-Computer Studies, 62(2), 263–298.

ROBERTS, S. T. (2019). Behind the screen: Content moderation in the shadows of social media. Yale University Press.

SCHWEMMER, C., & ZIEWITZ, M. (2022). Moderation at scale: Content moderation, collective control, and the regulation of platforms. Social Media + Society, 8(1). https://doi.org/10.1177/20563051221082720

SUZOR, N., WEST, S. M., QUODLING, A., & YORK, J. C. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. International Journal of Communication, 13, 1526–1543. https://ijoc.org/index.php/ijoc/article/view/9736
https://www.nbcnews.com/tech/tech-news/big-tech-companies-reveal-trust-safety-cuts-disclosures-senate-judicia-rcna145435
https://techxplore.com/news/2024-03-google-trims-jobs-safety-clock.html