Authentification biométrique : comment (ré)concilier sécurité, utilisabilité et respect de la vie privée ?

Estelle Cherrier

Normandie Université

Habilitation à Diriger des Recherches

Pour obtenir le diplôme de d'habilitation à diriger des recherches

Spécialité INFORMATIQUE

Préparée au sein de l'ENSICAEN et de l'UNICAEN

Authentification biométrique : comment (ré)concilier sécurité, utilisabilité et respect de la vie privée ?

Présentée et soutenue par
Estelle CHERRIER

HDR soutenue publiquement le 6 juillet 2021 devant le jury composé de		
Maryline LAURENT	Professeur, SAMOVAR, Telecom SudParis	Rapporteur
Jean-Luc DUGELAY	Professeur, EURECOM, Sophia-Antipolis	Rapporteur
Amine NAIT-ALI	Professeur, LISSI, UPEC	Rapporteur
Caroline FONTAINE	DR CNRS, LSV, Paris-Saclay	Examinatrice
Christophe ROSENBERGER	Professeur, GREYC, ENSICAEN	Examinateur
Christophe CHARRIER	Maître de Conférences HDR, GREYC, Normandie Université, UNICAEN	Examinateur (garant)

UNIVERSITÉ CAEN NORMANDIE

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

GREYC

Authentification biométrique : comment (ré)concilier sécurité, utilisabilité et respect de la vie privée ?

Estelle Cherrier

À Fabrice,
Lucas et Mathias

Remerciements

Remercier les personnes qui ont été proches de moi ces dernières années et qui ont donc contribué, directement ou indirectement, à l'aboutissement de mon projet d'HDR est un passage obligé de tout manuscrit, qui n'est pas le plus simple à écrire.
Voici donc quelques remerciements, en commençant par les membres du jury.
Merci à Maryline Laurent, Amine Nait-Ali et Jean-Luc Dugelay d'avoir accepté de rapporter ce manuscrit. Merci pour leurs remarques pertinentes et leurs appréciations chaleureuses et constructives.
Merci à Caroline Fontaine de s'être laissé convaincre de participer au jury de cette HDR.
Merci à Christophe Rosenberger d'être membre du jury et de m'avoir accueillie dans l'équipe SAFE (Monétique & Biométrie à l'époque) en 2011. J'y ai découvert une autre façon de faire de la recherche, ainsi qu'une ambiance chaleureuse, et un humour... inqualifiable – au sens premier du terme !
Merci à Christophe Charrier d'avoir accepté la mission importante et néanmoins symbolique de garant d'HDR. Merci pour nos conversations scientifiques et moins scientifiques, merci pour ton soutien, ton humanité et ta confiance.
Merci à mes collègues du GREYC ou de l'ENSICAEN. Merci à mes collègues de bureau : merci à Sylvain, qui est devenu mon super-binôme agile depuis un an ; merci à Baptiste, qui m'a initiée au véritable espresso avant de partir voir si l'herbe était plus orange chez le voisin ; merci à Patrick, qui m'a accueillie immédiatement. Merci à l'équipe des Monéticiens : Ndiaga, Alex, Joan, Benoît, Julien, Wilfried, d'avoir fait une place à la novice que je suis. Merci aux autres collègues de l'équipe SAFE pour nos cafés et discussions.
Merci à mes doctorant(e)s : Safa, Abir, Julien, Syed et Rima, vous m'avez accompagnée ces dix dernières années, ce manuscrit est un peu le vôtre !
Merci à Frédéric Jurie et Jalal Fadili d'avoir trouvé les arguments pour achever de me convaincre que j'étais prête.
Merci à Samia qui m'a prodigué des conseils d'amie pour une bonne HDR.
Merci à mes amis de l'Ensicaen : Cécile, Hugo, Christelle, Hélène. Merci pour votre soutien, votre présence, vos sourires.
Merci à mes parents, Marie-Claude et Joël, ma soeur Mylène et toute sa famille, Alex, Tristan et Emma, mes grands-parents, Huberte et Jean : même si vous êtes loin, vous êtes mes racines, vous êtes avec moi.
Et enfin, merci à ma famille, mes trois hommes : merci Lucas et Mathias, j'admire les hommes que vous êtes (presque) devenus. Lucas est né au début de ma thèse, il y a 18 ans, Mathias à la fin, il y a 15 ans. Vous étiez déjà là à ma soutenance de thèse, je compte sur votre soutien actif à ma soutenance d'HDR – au moins au moment du pot !
Comme le veut la coutume, le meilleur pour la fin. Merci à Fabrice, pour tout. Vraiment pour tout.

Table des matières

Table des figures

Préambule

L'esprit cherche et c'est le cœur qui trouve

George Sand

« Il y a vingt ans, le petit bourgeois français refusait de laisser prendre ses empreintes digitales, formalité jusqu'alors réservée aux forçats [...]
Au petit bourgeois français refusant de laisser prendre ses empreintes digitales, l'intellectuel de profession, le parasite intellectuel, toujours complice du pouvoir, même quand il paraît le combattre, ripostait avec dédain que ce préjugé contre la Science risquait de mettre obstacle à une admirable réforme des méthodes d'identification, qu'on ne pouvait sacrifier le Progrès à la crainte ridicule de se salir les doigts.
Erreur profonde ! Ce n'était pas ses doigts que le petit bourgeois français [...] craignait de salir, c'était sa dignité, c'était son âme.
Mais tout en se félicitant de voir la Justice tirer parti, contre les récidivistes, de la nouvelle méthode, il pressentait qu'une arme si perfectionnée, aux mains de l'État, ne resterait pas longtemps inoffensive pour les simples citoyens. »[1]

Ce texte de Georges Bernanos, écrit en 1947, cristallise les enjeux de sécurité et de respect de la vie privée inhérents à la biométrie. La « nouvelle méthode » dont il est question ici fait référence à la méthode Bertillon, qui se développe dans la police et le système judiciaire français depuis le début du XX$^{\text{ème}}$siècle. La biométrie fascine, elle est la preuve du caractère unique de chaque être humain, et elle effraie en même temps, par les dérives sécuritaires qu'elle laisse entrevoir. Si cette ambivalence a perduré à travers les dystopies, d'abord dans la littérature d'anticipation puis au cinéma, aujourd'hui l'argument de la facilité d'usage semble prédominer.
Cette dichotomie est de plus en plus présente dans la société actuelle, entre les besoins de sécurité et le respect de la vie privée, besoins enrichis par les exigences d'utilisabilité des systèmes biométriques. Ces trois enjeux constituent les trois axes de ma réflexion sur la biométrie et seront déclinés tout au long de ce manuscrit.

1. Extraits de *La France contre les robots*, Georges Bernanos, 1947

J'ai été nommée Maître de conférences à l'ENSICAEN, avec un rattachement au laboratoire GREYC, en septembre 2007. Mes thématiques de recherche concernaient l'Automatique, plus précisément la synthèse d'observateurs pour la synchronisation de systèmes chaotiques. J'ai découvert la biométrie il y a tout juste dix ans, lorsque Christophe Rosenberger m'a proposé d'encadrer la thèse de Rima Belguechi. Le sujet de Rima portait sur les techniques de biométrie révocable, qui garantissent le respect de la vie privée des utilisateurs. J'ai décidé de construire un nouveau projet de recherche, et j'ai quitté l'équipe Automatique du GREYC pour intégrer l'équipe Monétique & Biométrie en janvier 2011. Depuis, j'ai encadré cinq thèses, dont une qui a débuté en février 2020. Mon *curriculum vitae* est détaillé dans l'annexe **??**.

Introduction

On résout les problèmes qu'on se pose et non les problèmes qui se posent.

Henri Poincaré

1 Généralités sur la biométrie

Selon le Larousse [2], *le vocable le plus exact pour décrire le champ de la biométrie serait sans doute celui d'*anthropométrie *(du grec anthropos, « homme », et metron, « mesure »)*.
Selon la CNIL (Commission Nationale de l'Informatique et des Libertés), la biométrie *regroupe l'ensemble des techniques informatiques permettant de reconnaître automatiquement un individu à partir de ses caractéristiques physiques, biologiques, voire comportementales*.

Un système biométrique est généralement composé de plusieurs phases, comme illustré à la figure 1 :

- **L'enrôlement**
  Il repose sur une phase d'extraction de caractéristiques (temporelles, fréquentielles, obtenues en combinant plusieurs données), et se termine par le stockage d'une référence – générée à partir de ces caractéristiques – liée à l'identité de l'utilisateur. Celui-ci est alors enregistré dans le système.
- **L'authentification ou l'identification**
  Il s'agit de deux phases où le système respectivement (i) vérifie l'identité revendiquée par l'individu (authentification), ou (ii) vérifie si l'individu fait partie des utilisateurs déjà enregistrés dans la base et autorisés (identification). Elles sont déclenchées suite à une requête.
  Dans ce manuscrit, seule l'authentification [3], aussi appelée vérification, est considérée.

2. https://www.larousse.fr/encyclopedie/divers/biométrie/27110
3. L'article 3-5 du règlement eIDAS définit l'authentification comme : *un processus électronique qui permet de confirmer l'identification électronique d'une personne physique ou morale, ou l'origine et l'intégrité d'une donnée sous*

La suite du processus consiste en une comparaison entre la référence et les caractéristiques nouvellement extraites, puis une prise de décision par le système : si les deux données sont suffisamment proches (selon un seuil prédéfini), l'utilisateur est accepté par le système (il est qualifié de *légitime* dans ce cas), sinon il est rejeté (il est alors qualifié d'*imposteur*). Le choix de ce seuil de décision (ou seuil de comparaison) est un critère fondamental et délicat de la conception d'un système biométrique.

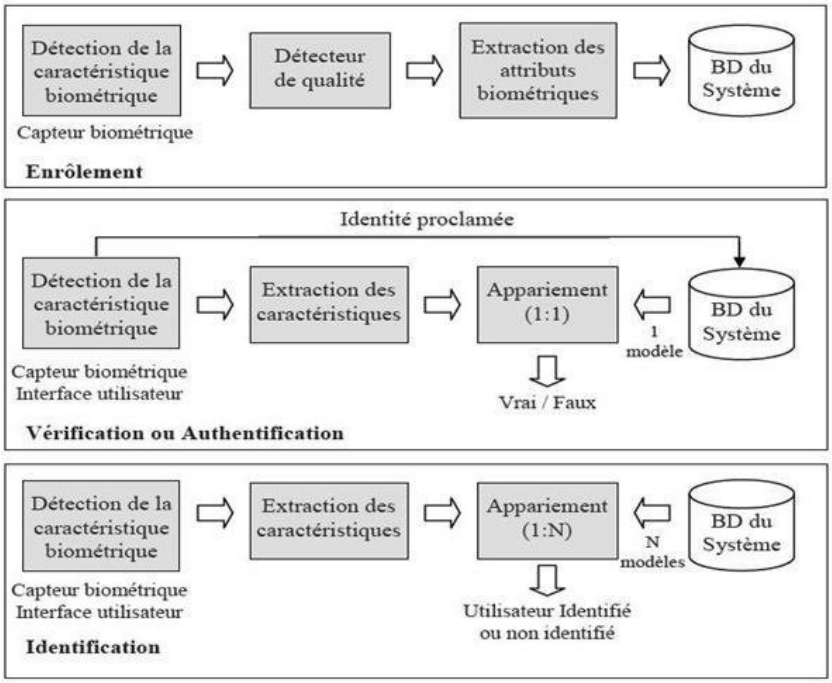


FIGURE 1 – Les étapes d'un système biométrique

A chaque étape (enrôlement ou vérification), on ajoute généralement **un pré-traitement des données** : le but est d'obtenir la référence (ou modèle, *template* en anglais) biométrique la plus représentative de l'utilisateur. Ce traitement fait appel à des outils de traitement du signal principalement pour « nettoyer » les données capturées, les normaliser si besoin, etc.

1.1 Modalité biométrique : définition et propriétés

Les caractéristiques précédentes constituent des *modalités* biométriques et sont au cœur des systèmes d'authentification. Au chapitre 1, les jalons historiques qui ont accompagné la découverte de ces modalités seront présentés, ainsi que l'évolution de leurs utilisations. Classiquement, les modalités biométriques sont regroupées en trois catégories :

- modalités *physiques* : empreinte digitale, visage, iris, forme de la main, voix, etc.,
- modalités *biologiques* : ADN, ECG, odeur, etc.,
- modalités *comportementales* : dynamique de frappe au clavier, dynamique de signature, interactions tactiles, voix, etc.

Ces modalités biométriques sont utilisées au sein de systèmes d'authentification. Elles constituent une catégorie parmi les facteurs d'authentification, définis comme suit :

forme électronique.

- **facteurs de connaissance** : il s'agit d'une information que l'utilisateur connaît, comme un mot de passe, un code PIN, un secret partagé ;
- **facteurs de possession** : il s'agit d'un élément, d'un objet que l'utilisateur possède, comme une carte à puce, une clé USB, un smartphone, un jeton de sécurité (ou *token*) ;
- **facteurs inhérents** ou **facteurs biométriques** : ce sont les seuls facteurs qui sont liés directement à l'utilisateur.

Pour être intégrée dans un système d'authentification biométrique, une modalité doit posséder un certain nombre de propriétés (voir par exemple les détails dans les articles [JDN04], [JRN11], [NF12]), parmi lesquelles :

- **Universalité** : détermine si la modalité existe et si elle est présente quel que soit l'individu ;
- Unicité : définit la probabilité de ne pas trouver de similarité entre les mesures d'une même modalité sur des individus différents (y compris sur des vrais jumeaux) ;
- **Permanence** : indique si la modalité reste présente et inchangée au fil du temps (hors accident ou vieillissement) ;
- **Facilité de collecte**, ou **collectabilité** : détermine le degré de facilité de l'acquisition, de la mesure et de l'exploitation de la modalité ;
- **Acceptabilité** : la collecte d'une modalité ne doit pas présenter un caractère intrusif pour l'utilisateur ;
- **Performance** : caractérise la robustesse, la fiabilité et la vitesse de la collecte de la modalité ;
- **Robustesse** ou **contournement** : représente la difficulté de contourner le système, par usurpation d'identité ou d'autres techniques de fraude

Le tableau 1 montre, pour différentes modalités, quel est le niveau (* bas, ** moyen, *** haut) atteint pour chaque propriété de la liste précédente. Ce tableau est une version mise à jour de celui publié dans la référence [BPJ98]. Il montre surtout qu'il ne peut exister de consensus autour d'une modalité particulière : chaque modalité privilégie une ou plusieurs propriétés parmi celles présentées ci-dessus, mais elles ne sont jamais toutes satisfaites en même temps. Cela implique une réflexion avant le choix d'une modalité pour la conception d'un système d'authentification biométrique. On peut cependant regretter que des modalités plus récentes, comme les interactions avec un écran tactile (schéma de déblocage d'un smartphone, données capturées par un smartphone, etc.) soient absentes de la comparaison. L'article très récent de Dargan et Kumar [DK20] présente un état de l'art des modalités biométriques les plus populaires.

1.2 Evolution de la problématique d'authentification

La problématique de la reconnaissance d'un individu a connu divers stades : autrefois, il s'agissait uniquement de reconnaître physiquement une personne présente dans un cercle restreint (famille, village) ; ensuite, les cartes d'identité, permis de conduire et autres passeports ont permis de déléguer le processus d'authentification à un tiers de confiance, pourvoyeur d'une identité administrative, légale. Aujourd'hui, la difficulté du défi s'est accrue : comment reconnaître *à distance* une personne *inconnue* ? Le fait que les facteurs biométriques soient les seuls facteurs d'authentification inhérents à un individu les rend extrêmement désirables, du point de vue des concepteurs et fournisseurs de services numériques principalement. Actuellement, les modalités retenues sont très variées. En premier lieu, on retrouve l'empreinte digitale et le

Modalité	Univ	Unic	Péren	Collec	Perf	Accept	Cont
Visage	***	*	**	***	*	***	*
Empreintes digitales	**	***	***	**	***	**	***
Paume de la main	**	***	***	**	***	**	***
Réseau vasculaire	**	**	**	**	**	**	***
Géométrie de la main	**	**	**	***	**	**	**
Rétine	***	***	**	*	***	*	***
Iris	***	***	***	**	***	*	***
Oreille	**	**	***	**	**	***	**
Démarche	**	*	*	***	*	***	**
Signature	*	*	*	***	*	***	*
Frappe au clavier	*	*	*	**	*	**	**
Voix	**	*	*	**	*	***	*
ADN	***	***	***	*	***	*	*

Tableau 1 – Propriétés des différentes modalités biométriques, inspiré de [BPJ98]

visage, modalités historiques toujours plébiscitées (*cf.* Face ID, Windows Hello, capteurs d'empreintes sur les nouveaux smartphones par exemple), en raison d'excellentes performances, exploitant les résultats de travaux de recherche obtenus sur plusieurs décennies. On peut citer d'autres modalités comme l'iris, la forme de la main, ou encore la voix, objets de recherches actives également. Les modalités comportementales sont en plein essor depuis quelques années, par exemple la dynamique de frappe au clavier, la dynamique de signature ou autres interactions avec un écran tactile, la démarche. Certains travaux commencent également à s'intéresser à des modalités dites « cachées », liées aux signaux biologiques comme l'EEG, l'ECG [GON20].

On constate aujourd'hui l'apparition de nouvelles catégorisations, par exemple les modalités biométriques *passives*[4], dans le sens où l'utilisateur – propriétaire d'un smartphone en l'occurrence – n'est pas actif dans la capture de ce type de modalités. Elles s'inscrivent dans un contexte d'authentification transparente (pour l'utilisateur) et continue (voir le chapitre 2), et appartiennent plus généralement au domaine de l'analyse comportementale (ou *behavioral analytics*), nouveau champ d'application de l'intelligence artificielle : le système apprend à reconnaître l'individu dans un contexte particulier. La collecte permanente d'informations provenant du smartphone d'un utilisateur (coordonnées GPS, données de l'accéléromètre, etc.) ainsi que ses données biométriques comportementales (voix, interactions tactiles, etc.) permettent de fluidifier certaines transactions (paiement sans contact ou en ligne, accès à des services bancaires, etc.) et par conséquent d'améliorer l'*expérience utilisateur*, comme nous le verrons au chapitre 2. Cependant, cette facilité d'usage ne doit pas obérer le respect de la vie privée des utilisateurs (*cf.* le chapitre 3), qui sont protégés depuis quelques années par le Règlement Général européen sur la Protection des Données.

4. https://mastercardcontentexchange.com/media/4322/evolution-of-biometrics_white-paper_v10.pdf, Mastercard, mai 2020

2 Cas d'usages et essor de la biométrie

Les cas d'usages de l'authentification biométrique se sont progressivement diversifiés, parmi lesquels on peut mentionner :

- respect des lois et sécurité publique (identification des suspects)
- domaine militaire (identification des ennemis)
- contrôle des frontières (authentification des voyageurs)
- services d'identité civile (authentification des citoyens, des contribuables, des électeurs)
- systèmes médical et social (authentification des bénéficiaires d'aides, des professionnels de santé à l'hôpital)
- accès physique et logique (authentification des employés, du propriétaire d'un ordinateur, d'un smartphone)
- applications commerciales (identification des consommateurs, authentification des clients), bancaires (paiement par reconnaissance de visage, d'empreinte, etc.)

Le premier facteur en faveur du déploiement des systèmes d'authentification biométrique vient du plus haut niveau de l'État. La figure 2 indique à quel point certains pays collectent et stockent les données biométriques de leurs citoyens, souvent au mépris de la notion même de respect de la vie privée : dans la légende de cette figure, plus le score est élevé, plus la collecte est importante. Cette collecte se fait notamment via les passeports, les visas et les contrôles aux frontières. Dans la liste des pays les plus intrusifs, on retrouve la Chine (score 24/25), puis la Malaisie et le Pakistan (score 21/25), ensuite les États-Unis (score 20/25) et l'Inde, l'Indonésie, les Philippines et Taïwan (score 19/25).

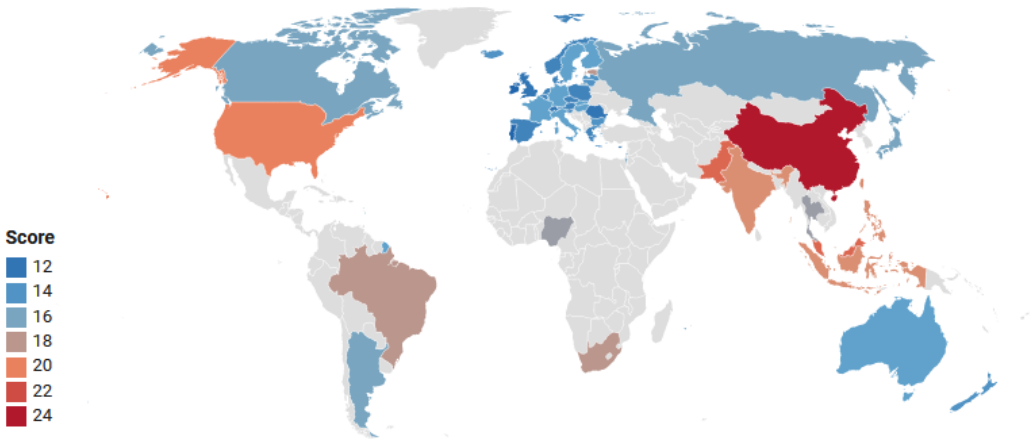


FIGURE 2 – Collecte et stockage de données biométriques par pays [5]

Par exemple, en Chine, la collecte de données est systématique et massive : numéro d'identification, origine ethnique, nationalité, numéro de téléphone, et même adresse, employeur, et des photos de chaque habitant, etc. Toutes ces informations sont intégrées dans un plan national appelé *Système de Crédit Social*, qui permet au gouvernent d'attribuer une note à chaque citoyen. Une sinologue explique ainsi dans les colonnes du Monde en janvier 2020 : « *L'idée est de collecter des centaines de données sur les individus et les entreprises, depuis leur capacité à tenir leurs engagements commerciaux jusqu'à leur comportement sur les réseaux sociaux, en passant par le respect du code de la route.* » [6] On peut souligner qu'en France, comme dans tous les pays de

5. https://www.comparitech.com/blog/vpn-privacy/biometric-data-study/, décembre 2019
6. https://www.lemonde.fr/idees/article/2020/01/16/le-credit-social-les-devoirs-avant-les-droits_6026047_3232.html

l'Union Européenne, un tel regroupement d'informations est interdit par le RGPD (Règlement Général sur la Protection des Données, *cf.* chapitre 3).

Les autres facteurs propices au déploiement des systèmes d'authentification biométrique, moins coercitifs, sont liés à (i) l'explosion des transactions commerciales (commerce électronique ou paiements sans contact principalement) et (ii) l'adoption d'un nombre grandissant d'objets dits *intelligents*. L'essor de ces services connectés a mis à mal le couple traditionnel {login + mot de passe} de l'authentification. En effet, il est de plus en plus difficile pour les utilisateurs de créer un mot de passe unique par service, d'autant plus que ce mot de passe doit être *suffisamment complexe*, donc compliqué à mémoriser. Malheureusement, de nombreux utilisateurs choisissent la facilité, soit en gardant le même mot de passe pour accéder à plusieurs services, soit en définissant un mot de passe trop simple, voire en adoptant une combinaison des deux comportements. L'enjeu de sécurité de ce mot de passe faible se trouve donc artificiellement augmenté, et ce, par la faute de l'utilisateur. Ainsi, en cas de compromission de ce mot de passe, les impacts sont nombreux : immédiats, comme l'usurpation d'identité, ou le vol de données, mais également à plus long terme si l'utilisateur, inconscient de cette compromission, ne le renouvelle pas. Une tendance actuelle est l'apparition sur le marché grand public de solutions matérielles (principalement sous la forme de clés USB) d'authentification à deux facteurs (Google Titan, Yubikey, OnlyKey, Thetis Fido U2F Security Key, etc.). Ces solutions sont compatibles avec les standards, principalement la norme d'authentification libre U2F – Universal Second Factor – gérée par la FIDO (Fast IDentity Online) Alliance. Quelques chiffres (études actuelles ou prévisions) attestent de ce véritable engouement pour l'authentification biométrique :

- le montant des transactions à base d'authentification biométrique sera de 45 milliards de dollars en 2024, suivant une croissance annuelle d'environ 20 %[7]
- en 2023, il y aura 1,5 milliard de smartphones recourant à la reconnaissance faciale [4]
- 42,5% des participants utilisent déjà leurs données biométriques pour un accès logique à leur smartphone, leur pc ou leur tablette, mais 58% préfèrent l'authentification par mot de passe uniquement [GB18]

Les figures 3 et 4, issues d'une étude de l'Université du Texas à Austin [GB18], qui comporte les réponses de 1000 individus, illustrent le ressenti actuel face aux systèmes biométriques. La modalité qui s'impose en termes de ressenti positif est l'empreinte digitale, sans doute grâce aux capteurs déjà présents sur la plupart des smartphones actuels. Quant aux raisons de la méfiance des utilisateurs, on peut constater que le non-respect de la vie privée arrive largement en tête, suivi par la crainte d'une surveillance étatique et d'une usurpation d'identité.
Ces deux figures illustrent bien le paradoxe : sous prétexte d'une meilleure sécurité et d'une simplification d'accès à des services connectés toujours plus nombreux, les systèmes d'authentification biométrique gagnent toujours plus de terrain ; cependant les utilisateurs sont tout à fait conscients des risques potentiels inhérents à ce type d'innovation et donc méfiants.

Par conséquent, cet essor des systèmes d'authentification biométriques impose une rigueur dans leur conception, ainsi qu'une sécurité garantie, gage d'acceptation par le grand public. S'y adjoignent des contraintes nouvelles en lien avec l'utilisabilité et le respect de la vie privée, dans le sens où les données biométriques font partie des données personnelles sensibles, selon le RGPD et la CNIL. La suite de cette introduction est consacrée aux différents tests et mesures qui permettent d'évaluer les systèmes biométriques, afin d'obtenir des éléments concrets et objectifs de comparaison, ou de discussion.
Pour un tour d'horizon des défis actuels auxquels la biométrie doit faire face, je renvoie vers les

7. https://www.biometricupdate.com/202003/global-biometrics-market-to-surpass-45b-by-2024-reports-frost-sullivan

	Very comfortable	Somewhat comfortable	Not very comfortable	Not at all comfortable at all
Eye recognition	34.30%	36.91%	19.56%	9.23%
Fingerprint scan	57.72%	28.36%	9.02%	4.91%
Voice recognition	36.47%	37.68%	17.64%	8.22%
Signature dynamics	38.68%	36.27%	17.94%	7.11%
Typing dynamics	36.07%	35.07%	20.24%	8.62%
Facial recognition	32.83%	36.75%	20.18%	10.24%
Hand geometry	40.42%	36.91%	16.95%	5.72%

FIGURE 3 – Degré de confiance face aux systèmes biométriques, en fonction des modalités [GB18]

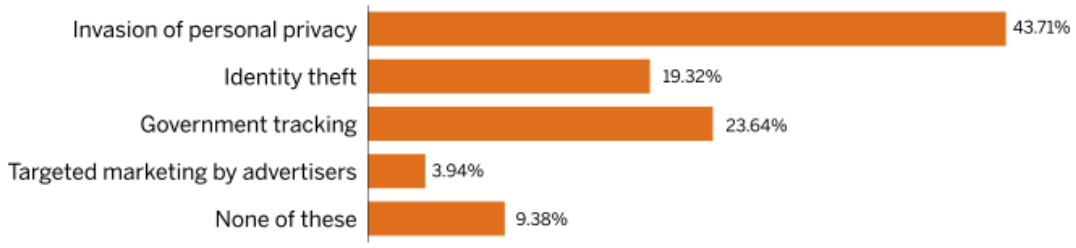


FIGURE 4 – Explications de la méfiance exprimée dans le tableau de la figure 3

deux articles récents [Akh+18] et [Ros+19].

3 Comment évaluer et tester un système biométrique ?

3.1 Composants d'un système biométrique

Dans ce qui suit, les principaux composants d'un système biométrique sont définis, en référence à l'illustration de la figure 5.

- **Sample, ou échantillon**
  Donnée biométrique présentée par l'utilisateur et capturée par le système (sous-système *Data collection*, *cf.* Figure 5) sous forme d'une image (empreinte digitale, iris, etc.), d'un signal (voix, dynamique de frappe au clavier, etc.)
- **Features, ou caractéristiques**
  Représentation mathématique des informations extraites de l'échantillon présenté par l'utilisateur. Cette étape se passe au niveau du sous-système *Signal Processing*. Ces caractéristiques sont soit stockées comme modèle dans la base de données (sous-système *Data Storage*, lors de la phase d'enrôlement), ou présentées au sous-système *Decision*, lors de la phase d'authentification, pour être comparées au modèle de référence correspondant à l'identité revendiquée par l'utilisateur.
- **Template, ou modèle de référence, ou gabarit**
  Une référence construite à partir des caractéristiques extraites de la donnée biométrique. Cette référence est associée à un identifiant unique de l'utilisateur.
- **Matching score, ou score de correspondance**

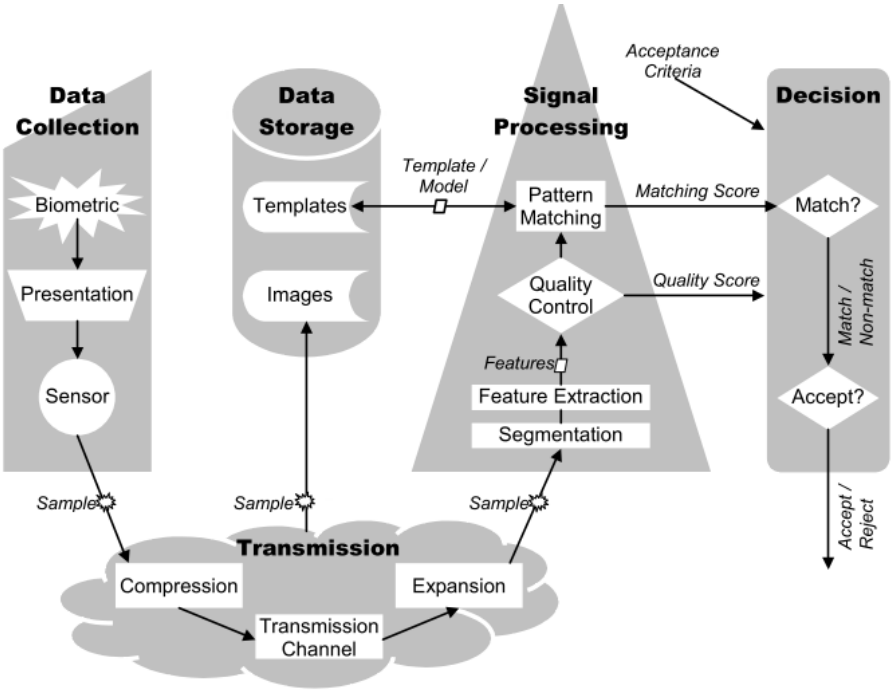


FIGURE 5 – Schéma d'un système biométrique, inspiré de l'ISO/IEC JTC1 SC37 SD11

Une mesure de la similarité ou de la distance entre deux gabarits : le premier correspond à l'identité clamée par l'utilisateur, le second est construit à partir de la donnée présentée au système par l'utilisateur.

- **Decision**
  A partir d'un seuil prédéfini, en fonction du score de correspondance, le sous-système *Decision* détermine si la donnée présentée par l'utilisateur est compatible (d'un point de vue statistique) avec la référence revendiquée.
- **Transaction**
  Tentative par un utilisateur de valider une revendication d'identité ou de non-identité (voir les définitions ci-dessous) : une transaction peut comporter une seule ou plusieurs présentations de la donnée biométrique au système, en fonction de la politique du sous-système de décision.

3.2 Différents types d'évaluations

Avant de déployer un système biométrique, il est nécessaire de le tester, de l'évaluer, afin de garantir qu'il remplit un certain nombre de contraintes. Dans la littérature, il est très courant de trouver des évaluations des performances techniques, reposant sur des taux de reconnaissance, des taux d'erreurs, etc., qui seront présentés juste après. Il existe également d'autres formes d'évaluations, comme détaillé dans l'article de Mansfield et Wayman [MW02]. Il est ainsi possible de tester :

- la fiabilité, la disponibilité, la maintenabilité
- la vulnérabilité
- la sécurité
- l'acceptabilité, l'utilisabilité
- les aspects humains
- le rapport coût/bénéfices

- le respect de la vie privée

Dans cette partie, je vais rappeler uniquement les outils nécessaires à l'évaluation des performances (techniques) d'un système biométrique, à savoir les outils statistiques de calcul de taux d'erreurs. Dans les chapitres suivants, je reviendrai sur l'évaluation de la sécurité, de l'utilisabilité et également du respect de la vie privée.

En fait, l'évaluation des performances se décompose en trois types, selon la série de standards ISO/IEC 19795 Biometric Performance Testing and Reporting (parties 1 à 7) et les articles [MW02], [Poh+11] :

- **Évaluation technologique**
  Le but de l'évaluation technologique est de comparer des algorithmes concurrents, pour une technologie donnée (un seul capteur) et une base de données fixée. Les performances vont dépendre de l'environnement et de la population testée. La règle dite des « Trois Ours » [MW02] s'applique : la base de données créée ne doit être ni trop difficile, ni trop facile. Dans cette configuration, les résultats sont reproductibles. Les tests sont réalisés *offline*.
- **Évaluation de scénario**
  Le but de l'évaluation de scénario est de déterminer la performance globale d'un prototype réaliste. Le capteur étant le seul élément qui change, un soin particulier est porté à la constitution de la base de données, ainsi qu'à l'environnement, qui doit rester le plus stable possible. Les tests reposent sur des comparaisons soit *online*, soit *offline* et sont reproductibles dans le cadre d'un environnement parfaitement contrôlé.
- **Évaluation opérationnelle**
  Le but de l'évaluation opérationnelle est de déterminer les performances d'un système biométrique dans un environnement particulier, avec une population cible particulière. Il s'ensuit que ces tests ne sont généralement pas reproductibles. Dans ces conditions, la vérité terrain peut être difficile à établir.

3.3 Les mesures de performances : erreurs de décision et erreurs de correspondance

On a vu précédemment que les systèmes biométriques permettent d'authentifier (ou d'identifier) un utilisateur. Cependant, on peut distinguer plusieurs acceptions pour l'authentification [MW02] :

1. **Revendication positive d'identité** (ou *positive claim of identity*)
   L'utilisateur prétend avoir été enrôlé dans le système. L'identité revendiquée peut prendre la forme d'un nom, un numéro d'identification personnel. C'est le cas pour les systèmes de contrôle d'accès la plupart du temps.
2. **Revendication négative d'identité** (ou *negative claim of identity*)
   L'utilisateur revendique ne pas être connu du système. On peut mentionner l'exemple des services sociaux réservés aux personnes qui ne sont pas déjà inscrites.
3. **Revendication explicite d'identité** (ou *explicit claim of identity*)
   Seule la référence correspondant à l'identité revendiquée est vérifiée. Il s'agit du **cas standard de l'authentification**, comportant une seule comparaison.
4. **Revendication implicite d'identité** (ou *implicit claim of identity*)
   Plusieurs comparaisons sont effectuées dans ce cas, avec revendication positive ou négative d'identité. Il s'agit du **cas standard d'identification**.

5. **Revendication légitime d'identité** (ou *genuine claim of identity*)
   L'utilisateur revendique son identité honnêtement : la comparaison est réalisée avec une référence correspondant réellement à cette identité.
6. **Revendication frauduleuse d'identité** (ou *impostor claim of identity*)
   L'utilisateur prétend faussement être quelqu'un d'autre, ce qui conduit à la comparaison avec un modèle qui ne correspond pas à l'identité revendiquée.

Pour rappel : on se place uniquement dans le cadre de l'authentification biométrique (ou vérification). Deux méthodologies peuvent être envisagées pour évaluer les performances du système : *a posteriori* ou *a priori*. L'évaluation *a posteriori* repose sur un unique ensemble de scores de correspondance, tandis que l'évaluation *a priori* requiert deux ensembles, à savoir un ensemble dit *de développement*, pour établir un seuil de décision en fonction d'un objectif à atteindre, et un second ensemble *d'évaluation*, pour calculer les performances du système, une fois le seuil fixé.
Sachant que l'unicité d'une donnée biométrique n'a jamais été mathématiquement démontrée, la biométrie repose sur des méthodes statistiques destinées à déterminer la probabilité que deux personnes présentent la même donnée. Même si les caractéristiques extraites étaient réellement uniques, les limitations ou autres imprécisions des techniques appliquées et des différentes circonstances de capture induiraient dans tous les cas une fiabilité partielle de l'authentification biométrique. La suite détaille les différents taux d'erreurs couramment utilisés.

Taux d'erreurs de décision

La majorité des performances des systèmes biométriques s'exprime en termes d'erreurs de décision, qui peuvent être de deux types :

- **False Accept Rate (FAR) ou Taux de fausse acceptation**
  Il s'agit de la proportion de transactions erronées, *i.e.* revendications d'identité par un imposteur dans le cas d'un système d'identité positif (cas n° 1 ci-dessus), ou de non-identité dans le cas n° 2, qui sont acceptées à tort. Dans le domaine mathématique, une fausse acceptation est souvent associée à une *erreur de Type II*, ou erreur de seconde espèce : l'hypothèse H_0 = « l'utilisateur est légitime » est fausse dans ce cas (il s'agit d'un imposteur), mais elle est acceptée à tort par le système.
- **False Reject Rate (FRR) ou Taux de faux rejet**
  Dans le cas d'un système d'identité positif (cas n° 1), il s'agit de la proportion de transactions (*i.e.* revendications d'identité dans le cas n° 1, ou de non-identité dans le cas n° 2) légitimes, qui sont faussement rejetées. Dans le domaine mathématique, un faux rejet correspond à une *erreur de Type I*, ou erreur de première espèce : l'hypothèse H_0 correspond cette fois à la vérité (l'utilisateur est bien légitime), mais elle est rejetée à tort par le système.

Taux d'erreurs de correspondance

On se place ici dans le cas d'une seule comparaison entre un échantillon présenté au système et un unique modèle de référence. Il faut noter que les taux de fausse acceptation ou de faux rejet (*i.e.* FAR ou FRR) sont des taux d'erreurs au niveau du système, qui prennent en compte les échantillons dont l'acquisition ou la comparaison a échoué. Les taux de fausse reconnaissance, ou de fausse non-correspondance définis ci-dessous se situent, quant à eux, au niveau algorithmique, et ne peuvent être calculés qu'*a posteriori*, lorsque toutes les données ont été capturées.

- **False Match Rate (FMR) ou Taux de fausse reconnaissance**
  Il s'agit d'une estimation de la probabilité que le système déclare à tort qu'un échantillon biométrique appartient à l'identité déclarée par l'utilisateur alors qu'il appartient en réalité à un sujet différent (dans ce cas l'utilisateur est un imposteur).
- **False Non-Match Rate (FNMR) ou Taux de fausse non-correspondance** Il s'agit d'une estimation de la probabilité que le système rejette à tort une identité clamée alors que l'échantillon appartient effectivement au sujet (dans ce cas l'utilisateur est légitime).

L'**EER** (Equal Error Rate) constitue l'indicateur privilégié de performance du système : il est défini comme le point (unique) où le FMR est égal au FNMR. En pratique, on calcule l'**HTER** (Half Total Error Rate), via la moyenne du FMR et du FNMR : l'EER empirique est approché par l'HTER au point où la différence entre le FMR et le FNMR est minimale.

Taux d'erreurs d'acquisition

- **Failure to enrol rate (FTE), ou taux d'échec à l'enrôlement**
  Ce taux correspond à la proportion de la population qui est dans l'incapacité à générer des échantillons reproductibles. Cela peut être dû à une donnée manquante, une image de qualité insuffisante.
- **Failure to acquire rate (FTA), ou taux de défaut d'acquisition**
  Ce taux est défini comme la proportion attendue de transactions pour lesquelles le système est incapable de capturer ou de localiser une image ou un signal de qualité suffisante. Il peut dépendre des seuils prédéfinis pour ajuster la qualité de l'image ou du signal.

Représentations graphiques des performances

Une courbe **ROC** (*i.e. Receiver Operating Characteristic curve*) est un graphique à deux dimensions qui illustre les performances du système. Il s'agit du tracé du taux de FNMR, ou de FRR (sur l'axe des ordonnées) en fonction du taux correspondant de FMR, ou de FAR (sur l'axe des abscisses), respectivement. La courbe est paramétrée par le seuil de décision fixé.

Une courbe **DET** (*Detection Error Trade-off curve*) est semblable à la courbe ROC, à l'exception de ses axes : ils suivent une échelle non-linéaire (logarithmique le plus souvent), afin de mettre en valeur certaines zones d'intérêt de la courbe.

3.4 Standardisation des technologies biométriques

La standardisation relève du domaine de l'*ISO/IEC JTC 1/SC 37 Biometrics*.
Elle consiste en la normalisation des technologies biométriques génériques ayant trait aux personnes en vue de prendre en charge l'interopérabilité et l'échange de données entre applications et systèmes. Les normes biométriques humaines génériques comprennent notamment [8] :

- les structures de fichiers communs
- les interfaces de programmation des applications biométriques
- les formats d'échanges de données biométriques
- l'application de critères d'évaluation aux technologies biométriques
- les méthodologies concernant les essais de performance

8. https://www.iso.org/fr/committee/313770.html

- les aspects juridictionnels et sociétaux

Le sous-comité 37 se répartit en six groupes de travail, ou *Working Group* : WG 1 - Vocabulaire ; WG 2 - Interfaces techniques ; WG 3 - Formats d'échange de données biométriques ; WG 4 - Mise en oeuvre technique des systèmes biométriques ; WG 5 - Méthodes de tests ; WG 6 - Aspects omnijuridictionnels et sociétaux de la biométrie.
A ces travaux s'ajoutent ceux du Sous-Comité (SC) 27, intitulé : *Sécurité de l'information, cybersécurité et protection de la vie privée*. Les normes établies par le SC 27 concernent les méthodes génériques, les techniques et les lignes directrices visant à traiter les aspects de sécurité et de protection de la vie privée [9]. Du point de vue de la biométrie, ces normes viennent compléter celles du SC 37. Parmi les enjeux, on peut citer les points suivants :
Ces enjeux de normalisation constituent actuellement de véritables verrous scientifiques, qui sous-tendent la quasi totalité des travaux de recherche en biométrie : si l'évaluation peut être réalisée pour un système particulier, elle peut également servir à comparer deux systèmes.

4 Organisation du manuscrit

Pour construire ce manuscrit, j'ai décidé de mettre l'accent sur trois thématiques essentielles en lien avec tout système d'authentification biométrique :

- la sécurité
- l'utilisabilité
- le respect de la vie privée et la protection des données personnelles

J'ai essayé de constituer un squelette commun aux trois chapitres. Après une courte introduction, le début de chaque chapitre est consacré à une prise de recul sur son thème central. Cette prise de recul peut mettre en jeu des éléments historiques, éthiques, juridiques, des normes et réglementations (européennes ou françaises). J'ai également voulu replacer chaque thème dans un contexte plus large, avec une discussion sur quelques points clés ou points de vue divergents. La seconde partie des trois chapitres détaille quelques contributions sous forme d'une synthèse de différents travaux issus des thèses que j'ai encadrées, synthèse suivie de quelques perspectives. Mon projet de recherche constitue le dernier chapitre. J'y développe des pistes de réflexion à court, moyen et long terme, qui devraient alimenter mes recherches pour les années à venir. Voici une rapide présentation de chaque chapitre.

Chapitre 1.
La sécurité est la raison d'être de l'authentification biométrique. Pour commencer ce chapitre, j'ai choisi de retracer l'évolution historique de la biométrie, depuis l'Antiquité et surtout le Moyen-Âge, en passant par la révolution apportée par le système Bertillon, jusqu'aux systèmes embarqués sur nos smartphones aujourd'hui. L'étude de la sécurité des systèmes biométriques comporte plusieurs niveaux. Après un passage obligé par les normes et contraintes réglementaires, le système biométrique est vu comme l'objet d'études de risques de sécurité. Différentes cartographies permettent de mettre à jour des vulnérabilités génériques. Les contributions sélectionnées se situent au niveau de l'amélioration des performances d'authentification (par dynamique de frappe au clavier), soit en ayant recours à la biométrie douce, soit en développant des techniques de mise à jour (ou d'adaptation) de modèle.

Chapitre 2.
L'utilisabilité est une thématique plus anecdotique que les deux autres dans mes travaux de recherche. Elle est éparpillée dans les différentes contributions, mais se trouve au cœur de mon

9. https://www.iso.org/fr/committee/45306.html

projet de recherche : la construction même de mon projet a placé l'utilisabilité au centre de mes réflexions. Si on oublie de prendre en compte l'utilisabilité, il est inutile de garantir un niveau de sécurité fantastique et un respect de la vie privée immense : le système d'authentification biométrique déployé ne sera pas validé par les utilisateurs, donc pas utilisé. L'utilisabilité oblige à la fois à mettre en œuvre des pratiques guidées par le bon sens, et à solliciter des ressorts psychologiques, du marketing, de la pédagogie, de l'UX design, etc.

Chapitre 3.
S'il ne devait rester qu'un seul chapitre, ce serait celui-ci. Quoique j'ai découvert des perspectives que je n'envisageais pas avant la rédaction de ce manuscrit concernant l'utilisabilité, que j'ai aujourd'hui envie d'explorer. La biométrie respectueuse de la vie privée, c'est la thématique qui m'a poussée à changer d'équipe et à laisser de côté l'automatique et la synchronisation des systèmes chaotiques. Découvrir que la protection des données biométriques est possible, mathématiquement, a été une révélation et je me suis lancée avec enthousiasme dans l'encadrement de la thèse de Rima Belguechi en juin 2010. Ce chapitre commence par différents points de vue (historique, éthique, et, bien sûr, informatique) sur le respect de la vie privée. S'ensuit une présentation de la réglementation en la matière, depuis le concept essentiel de *Privacy by Design*, jusqu'au RGPD. La dernière partie sur la biométrie révocable, est centrée sur le BioHashing, algorithme populaire de protection des données biométriques, au cœur d'un système d'authentification à deux facteurs, et au cœur de la thèse de Rima.

Chapitre 4.
J'ai commencé à réfléchir à ce projet il y a deux ans. Sa mise en forme m'a confortée dans l'idée de rédiger cette HDR. Il prend une tournure plus précise depuis septembre dernier, avec le lancement d'un certains nombre d'actions concrètes. Ce projet s'ouvre sur pistes à court/moyen terme – avec pour thématique principale la collecte de données comportementales respectueuse de la vie privée, centrée sur l'utilisateur –, et sur des pistes à plus long terme – concernant la modélisation des données biométriques, ou leur protection.

Remarque : j'ai choisi de conserver certaines formulations ou expressions en anglais, lorsque la traduction française n'existe pas, n'est pas usitée, est peu élégante. Dans ce cas, les expressions sont écrites en italique.

Références de l'Introduction

[Akh+18] Z. AKHTAR, A. HADID, M. S. NIXON, M. TISTARELLI, J. DUGELAY et S. MARCEL. « Biometrics: In Search of Identity and Security (Q & A) ». In : *IEEE MultiMedia* (2018).

[BPJ98] R. BOLLE, S. PANKANTI et A. K. JAIN. *Biometrics, Personal Identification in Networked Society*. Kluwer Academic Publishers, 1998.

[DK20] S. DARGAN et M. KUMAR. « A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities ». In : *Expert Systems with Applications* 143 (2020), p. 113114.

[GB18] R. L. GERMAN et K. S. BARBER. *Consumer Attitudes About Biometric Authentication*. Rapp. tech. Center for Identity - The University of Texas at Austin, 2018.

[GON20] B. GOUDIABY, A. OTHMANI et A. NAIT-ALI. « EEG Biometrics for Person Verification ». In : *Hidden Biometrics: When Biometric Security Meets Biomedical Engineering*. Sous la dir. d'A. NAIT-ALI. Springer Singapore, 2020.

[JRN11] A. K. JAIN, A. A. ROSS et NANDAKUMAR. *Introduction to biometrics*. Springer Publishing Company, Incorporated, 2011.

[JDN04] A. K. JAIN, S. C. DASS et K. NANDAKUMAR. « Soft Biometric Traits for Personal Recognition Systems ». In : *Biometric Authentication*. Sous la dir. de D. ZHANG et A. K. JAIN. Berlin, Heidelberg : Springer Berlin Heidelberg, 2004, p. 731-738.

[MW02] A. J. MANSFIELD et J. L. WAYMAN. *Best Practices in Testing and Reporting Performance of Biometric Devices*. Rapp. tech. Centre for Mathematics et Scientific Computing, 2002.

[NF12] A. NAIT-ALI et R. FOURNIER. *Traitement du signal et de l'image pour la biométrie*. Hermes/Lavoisier, 2012.

[Poh+11] N. POH, C.-H. CHAN, J. KITTLER, J. FIERREZ et J. GALBALLY. *Description of Metrics For the Evaluation of Biometric Performance*. BEAT Project (Biometrics Evaluation and Testing). 2011.

[Ros+19] A. ROSS et al. « Some Research Problems in Biometrics: The Future Beckons ». In : *12th IAPR International Conference on Biometrics (ICB), Crete, Greece*. 2019.

CHAPITRE

1

Biométrie et sécurité

Do not figure on opponents not attacking ; worry about your own lack of preparation

The Book of Five Rings, Miyamoto Musashi, 1644

Biométrie et sécurité. La juxtaposition de ces deux termes renvoie à différents contextes, différents niveaux, chacun conférant à cette expression un sens particulier.
Ce chapitre est scindé en trois parties. La première apporte une vision assez large de la sécurité des systèmes biométriques. Elle commence par des repères historiques de l'évolution conjointe des deux domaines, celui de la sécurité et celui de la biométrie. La deuxième partie est consacrée aux études de sécurité des systèmes biométriques. Les normes et contraintes réglementaires traitant de la sécurité des systèmes d'information et des technologies de l'information sont rapidement présentées. Y sont associées les traditionnelles méthodes d'évaluation et d'analyse des risques, via des cartographies de vulnérabilités et autres modèles de menaces. Cette partie se termine par un focus sur la sécurité des données biométriques. Dans la troisième partie, je présente des contributions issues des thèses que j'ai encadrées, qui sont en lien avec ces aspects de sécurité dans les systèmes biométriques. Il s'agit plus précisément de deux thèses, celle de Syed Idrus Syed Zulkarnain, focalisée sur la biométrie douce pour la dynamique de frappe au clavier [Sye14] et celle d'Abir Mhenni, qui s'est intéressée à la mise à jour du modèle biométrique, aussi appelée adaptation de ce modèle [Mhe19].

1 Biométrie et sécurité : quelques éléments historiques

Pour pouvoir étudier les liens entre biométrie et sécurité, il semble intéressant de considérer ces deux notions d'un point de vue historique, car elles ont été étroitement liées dès les débuts de la biométrie.
On peut distinguer plusieurs phases dans l'évolution de la biométrie depuis le Moyen-Âge [Abo11] :

- du Moyen-Âge jusqu'au XVIII$^{\text{ème}}$siècle : *préhistoire* de la biométrie
- XVIII$^{\text{ème}}$siècle : les progrès de l'anthropométrie

- XIX$^{\text{ème}}$siècle – milieu du XX$^{\text{ème}}$siècle : développement d'une science de l'identification
- début XX$^{\text{ème}}$siècle : le système Bertillon
- à partir des années 1970 : systèmes automatiques de reconnaissance biométrique

1.1 La préhistoire de la biométrie

Si on met à part les empreintes palmaires, déjà utilisées par les scribes égyptiens (aux environs de -3000 av. JC) à des fins d'identification, ou par les chinois (paume de main encrée, au III$^{\text{ème}}$siècle av. JC), pour authentifier des documents, la biométrie, en tant que science de l'identification, n'a pas connu de grands progrès jusqu'au Moyen-Âge. A cette époque, il s'agit simplement de reconnaître des personnes recherchées par la Justice, par le biais de descriptions de traits corporels, de la taille, l'apparence, etc. La mise en place de registres pour transmission entre services éloignés géographiquement nécessite le développement de l'écriture, puisque les personnes recherchées voyagent à travers le pays. Cela explique pourquoi la reconnaissance biométrique n'a pas été matériellement possible auparavant.
Avec le déploiement des relations postales, combinées à une intense mobilité en Europe, le XVI$^{\text{ème}}$siècle voit une augmentation des procédures de transmission des signalements de fugitifs, principalement des criminels et déserteurs, puis des esclaves au XVIII$^{\text{ème}}$siècle. Ces signalements se fondent sur les signes sur le corps, ou *signa* : taches de naissance, grains de beauté, cicatrices, etc., qui deviennent des preuves, ou *evidentia* entre les mains des magistrats. On considère d'ailleurs à cette époque que ces différentes marques sur le corps sont *le reflet de l'âme des criminels* [Den11].
En 1678-1679, Nehemiah Grew publie *Philosophical Transactions*, un traité détaillé sur les empreintes digitales – plus précisément sur les dermatoglyphes – destiné à la *Royal Society*, qui mentionne leurs « innombrables petites rides ».

1.2 Les progrès de l'anthropométrie

Au XVIII$^{\text{ème}}$siècle, le domaine de l'anthropométrie progresse, grâce à la définition de nouvelles caractéristiques, telles que l'âge, la corpulence, les cheveux, le visage, ainsi qu'une description du corps de haut en bas. Une première tentative de mesure objective de la taille donne lieu à une expérimentation dans les hôpitaux : faute de moyens, aucune suite n'est envisagée.
La définition d'un « signalement » fait son apparition dans le Dictionnaire de l'Académie en 1718 :

Un signalement est la description que l'on fait de la figure d'un déserteur ou d'un criminel, et que l'on donne pour le faire reconnaître.

En parallèle, on peut noter l'apparition des premiers papiers d'identité. On peut mentionner l'appartenance à un corps de métier chez les compagnons : par exemple, les ouvriers se déplacent sur des chantiers éloignés et peuvent prouver qu'ils ne sont pas des vagabonds recherchés par la Justice. C'est aussi à cette époque qu'apparaissent les premiers passeports, qui vont pouvoir combler les besoins de centralisation de l'État. Le passeport sera rendu obligatoire pour voyager par Napoléon en 1807.
Même si la description des individus a progressé, de nombreux obstacles subsistent : la description est inévitablement subjective, n'étant encadrée par aucun processus de normalisation ; l'autre problème qui se pose est celui de la transmission d'informations : les registres doivent être constamment mis à jour et dans le même temps partagés avec les juridictions de tout le pays, ce qui est impossible à réaliser à cette époque. Les progrès vont venir de deux domaines : les statistiques et les techniques de communication.

1.3 Naissance d'une Science de l'identification

La problématique au XIX^{ème}siècle vient d'un nombre trop important de signalements (suite à des évasions par exemple), à mettre en regard des possibilités de classement qui sont, elles, insuffisantes. L'impulsion, sous Napoléon, d'une uniformisation judiciaire et administrative de la France requiert la mise en relation des caractéristiques physiques avec l'identité civile. Un début de solution apparaît, d'une part avec l'élaboration d'un dictionnaire permettant de mettre en place un vocabulaire commun en 1830, puis d'un registre des marques distinctives en Grande-Bretagne en 1870. Ce registre, en suivant l'ordre alphabétique, propose de parcourir la surface du corps, qui est littéralement quadrillée. L'identification se développe. Par la suite, les informations s'accumulent dans les registres, qui deviennent rapidement obsolètes et doivent être réécrits très régulièrement. Autour des années 1880-1900, dans plusieurs domaines en lien avec la nouvelle science de l'identification, on retrouve des métiers en pleine expansion : expert policiers (des services policiers se spécialisent dans tous les pays), magistrats, anthropologues, médecins, etc. C'est également à cette époque que le médecin italien Lombroso développe sa théorie du *criminel-né*, en lien avec le poids du cerveau.
Une véritable science des indices voit le jour, à partir de la photographie judiciaire, la mesure de parties du corps, le classement des empreintes digitales. En 1823, le tchèque J. Purkinje classe les empreintes en 9 motifs. En 1860, le britannique W. Herschel constate que les empreintes (digitales) sont « *formées avant la naissance et restent inchangées* ». Cela lui permet d'utiliser ses propres empreintes pour signer des contrats, ou des chèques. En 1872, J. Bonomi, spécialiste de la statuaire ancienne dans l'Antiquité orientale établit une formule, suffisamment précise, pour répartir les fiches dans des sections de taille inférieure. Cette formule définit un indice de proportion entre la taille et l'envergure des bras (comme dans la représentation de l'Homme de Vitruve). Il s'appuie sur l'Art, la Science et la Justice. En 1884, F. Galton propose un système de cartes perforées (appelé *Mechanical Selector*), permettant de réaliser un classement simple des fiches en fonction des mesures du corps : il supprime de fait l'étape (subjective) d'interprétation des fiches, par un processus mécanique. Par ailleurs, F. Galton et E. Henry travaillent sur la classification des empreintes digitales. Une empreinte digitale est divisée en quatre zones : la zone centrale (le centre de l'image), la zone basale (la partie basse), la zone distale (la partie haute) et les zones marginales (les côtés). Ensuite il faut s'intéresser à la présence de deltas : il s'agit d'un point de convergence entre les zones centrale, basale et marginales, en forme de triangle ouvert ou fermé, ou formé directement par une crête [1].
D'un autre côté, grâce aux progrès de l'anthropométrie, de nouvelles techniques de classement apparaissent. Les données sont réparties dans trois fichiers : un fichier alphabétique, un fichier anthropométrique (en 1880), et un fichier dactyloscopique (en 1890). A partir de là, les empreintes digitales deviennent un enjeu scientifique majeur, qui trouvera son aboutissement dans le système Bertillon.

1.4 Le système Bertillon

Le criminologue français, Alphonse Bertillon (1853-1914), est le véritable créateur de l'anthropométrie judiciaire, qui donnera naissance à la reconnaissance biométrique quelques années plus tard [2]. Il s'appuie sur les progrès dans deux domaines : l'anthropologie et la statistique. L'essor des connaissances en anthropologie en France a lieu sous l'impulsion de P. Broca, qui fonde en 1859 la Société d'anthropologie de Paris, puis l'École d'anthropologie de Paris en 1876.

1. https://www.police-scientifique.com/empreintes-digitales/type-de-dessin-et-classification/
2. Voir les dossiers du site https://criminocorpus.org/fr/bibliotheque/collections/police-scientifique-bertillonnage/

Par ailleurs, les progrès de la statistique appliquée aux proportions du corps proviennent des travaux du belge A. Quételet au XIX$^{\text{ème}}$siècle. Il applique des concepts de l'astronomie à la société : son but est de comprendre si les phénomènes humains présentent les mêmes irrégularités que les phénomènes naturels. Il est le premier à établir la moyenne d'une mesure du corps humain, en l'occurrence le tour de poitrine de soldats écossais. Puis il va calculer la moyenne de toutes les mesures possibles du corps humain. De là naît le concept plus que controversé d'« *homme moyen* », « *moyen* » étant pris dans une acception étonnante. En effet, pour Quételet, ce terme désigne l'homme idéal, parfait, sans défaut :

> « *Si l'homme moyen était parfaitement déterminé, on pourrait, comme je l'ai fait observer déjà, le considérer comme le type du beau ; et tout ce qui s'éloignerait le plus de ressembler à ses proportions ou à sa manière d'être constituerait les difformités et les maladies ; ce qui serait dissemblable, non seulement sous le rapport des proportions et de la forme, mais ce qui sortirait encore des limites observées, serait monstruosité.* » [3]

Via un glissement des considérations physiques vers les considérations morales, le criminel est vu à cette époque comme une exception sociale et biologique, dans le sens où il se situe loin du comportement de l'*homme moyen*.

De par sa position à la préfecture de Police de Paris (il y classe les fiches signalétiques), Alphonse Bertillon constate les failles et lacunes du système de classement : certaines ressemblances physiques génèrent une concentration indésirable des fiches dans certaines zones du fichier. En outre, certaines mesures se situent à la limite, et induisent un risque de confusion et/ou des erreurs de triage. Les améliorations proposées par Bertillon s'articulent dans quatre domaines [Abo11]. L'*anthropométrie*, enrichie par des alternatives nouvelles comme l'oreille (à partir des travaux de Lannois et Frigerio), le nez ou l'œil (iris, contour, dispositions des veinules, etc.) ; le *portrait parlé* du visage et du corps, constituant une description très détaillée ; le *portrait photographique*, comportant des vues (devenant de plus en plus standardisées) de face et de profil ; le *relevé des marques particulières* des mains, du corps, recensant et localisant tous les signes corporels (professionnels), cicatrices, tatouages, grains de beauté, etc.

Les améliorations proposées par Bertillon constituent donc ce qu'on appelle le « *portrait parlé* » : des photographies (du visage) et un signalement descriptif. L'identification des récidivistes devient un enjeu scientifique majeur au tournant du XX$^{\text{ème}}$siècle. Les hommes politiques et le grand public se saisissent du sujet, les connaissances se diffusent. Tout cela concourt à la naissance de l'identité judiciaire moderne, dont la référence mondiale est le système Bertillon.

Par la suite, Bertillon va enrichir son système, dont la principale limitation est son incapacité à apporter une preuve indiscutable de l'identité d'une personne. Différentes parties du corps sont étudiées et strictement cartographiées, à la manière des leçons d'anatomies de l'époque : l'arc veineux du dos de la main (en suivant les travaux d'A. Tamassia, en 1908) est décalqué, photographié et réparti entre six catégories ; l'oreille, décrite par M. Lannois en 1887 et L. Frigerio en 1888 ; l'œil est particulièrement détaillé (son contour, la couleur et la description de l'iris, le fond de l'œil, la disposition des veinules, etc.), à partir des travaux de P. Broca, en lien avec la place accordée au regard dans l'étude de la déviance, dans la théorie de l'*anthropologie des races*. Grâce à la collecte de ces différentes connaissances, Bertillon atteint son objectif de définir un vocabulaire précis pour décrire les personnes interpellées par la police. Il donne même son nom au *bertillonnage*, qui parvient au niveau d'une véritable *science criminelle* en permettant de distinguer deux individus distincts. Ce sont les empreintes digitales qui vont permettre de réellement résoudre des crimes, comme lors de l'affaire Scheffer, en 1902 : Bertillon parvient à confondre un assassin déjà fiché grâce à ses empreintes. C'est le début du déclin de l'anthropométrie au profit de la dactyloscopie : ainsi, en 1907, l'Académie française des sciences publie

3. *Sur l'homme et le développement de ses facultés, essai d'une physique sociale*, Tome second, A. Quételet, 1835

un rapport dans lequel elle reconnaît la supériorité de la seconde sur la première en matière d'identification.

1.5 Systèmes automatiques d'identification et de reconnaissance biométrique

Le but ultime de toutes ces recherches est ensuite de réduire le corps à une série de chiffres, appelée *code d'identité*, ceci afin de pouvoir transmettre les signalements sous forme de codes courts, reconnus internationalement, grâce aux technologies émergentes de l'époque : le télégraphe, le téléphone. C'est chose faite avec les travaux de S. Icard [Ica09], qui donne l'impulsion d'une union judiciaire internationale dès 1908 : il définit la *fiche-numéro dactylo-anthropométrique*, composée de vingt-et-un chiffres. Les dix premiers chiffres correspondent à des données de dactyloscopie, les onze chiffres restants correspondent à une formule anthropométrique qui décrit onze mesures du corps par un chiffre égal à 1, 2 ou 3 (*i.e. petit*, *moyen*, *grand*). En pratique, le registre international des fiches-numéros est inutilisable, car il recense un million d'individus dans six volumes de mille pages. L'étape suivante repose sur la dématérialisation des photographies (portrait ou empreintes digitales) sous forme de codes constitués de lettres et chiffres. Les deux Guerres Mondiales marquent un ralentissement dans la propositions de nouvelles techniques. Dans les années 1970, le couplage des techniques classiques d'identification biométrique avec l'informatique permet des avancées significatives, notamment l'automatisation des processus. Ce progrès est crucial, car jusqu'alors, l'identification pouvait être réalisée par : un opérateur dédié à la reconnaissance visuelle des individus ; une possession (badge) ; une connaissance (code secret, mot de passe, etc.). Aujourd'hui, les systèmes biométriques sont des systèmes automatisés permettant de reconnaître les individus avec une grande précision sans avoir besoin d'opérateur dédié, de carte ni de mot de passe.
On constate ainsi que la problématique liée à la biométrie a longtemps été l'identification. L'authentification biométrique est un usage assez récent en regard de l'historique précédent, lié à une demande de facilité d'utilisation (*cf.* le chapitre 2) dans l'accès à certains équipements personnels principalement.
Parce qu'ils traitent des données sensibles, qui plus est non révocables (*cf.* le chapitre 3), les systèmes biométriques sont l'objet d'études de sécurité à plusieurs niveaux.

2 La sécurité des systèmes d'authentification biométrique

Comme présenté par Brömme dans l'article [Brö06], un système d'authentification biométrique peut être considéré comme un élément d'une architecture générale sécurisée d'authentification. C'est dans ce sens que l'étude de la sécurité d'un système biométrique est incluse dans le contexte général de l'évaluation de la sécurité d'un Système d'Information (SI), dans le domaine global des Technologies de l'Information et de la Communication (TIC). La sécurité est ainsi analysée selon une approche holistique, et pas simplement comme une étude de type boîte noire, centrée uniquement sur les algorithmes d'enrôlement et de vérification.
Ainsi l'étude de la sécurité des systèmes d'authentification biométrique peut être menée selon trois axes, du plus général au plus spécifique :

- la sécurité des SI
- la sécurité des systèmes biométriques
- la sécurité des données biométriques

D'autres approches présentent ces trois niveaux d'évaluation de la manière suivante :

- des évaluations qualitatives, reposant par exemple sur les Critères Communs, les normes ISO, les méthodes d'analyse de risques ;

- des évaluation quantitatives, c'est-à-dire à base de modèles (de vulnérabilités, de menaces), impliquant une évaluation conjointe des compétences de l'attaquant, des efforts à fournir pour exploiter les vulnérabilités identifiées dans le système ;
- des évaluations expérimentales, reposant sur des tests de performances, sur des données réelles.

2.1 Normes et contraintes réglementaires

On retrouve ces distinctions dans les différentes normes qui traitent de la sécurité des SI (ou des TIC) et de la biométrie. Toutes relèvent du JTC1 (Joint Technical Committee) de l'ISO/IEC (International Organization for Standardization/International Electrotechnical Commission), dont le thème est : les technologies de l'information. Les deux sous-comités SC27 et SC 37, déjà mentionnés à la page 11, s'intéressent particulièrement à la sécurité des systèmes biométriques.

- Le sous-comité ISO/IEC JTC1-SC27 *Sécurité de l'information, cybersécurité et protection de la vie privée* [4].
  Parmi les normes concernant les méthodes génériques, les techniques et les lignes directrices visant à traiter les aspects de sécurité et de protection de la vie privée, on peut mentionner : *les critères et la méthodologie d'évaluation de la sécurité*, ou encore *les aspects de sécurité de la gestion des identités, de la biométrie et de la protection de la vie privée*.
  Comme tous les sous-comités, l'un des objectifs du SC27 est d'établir des normes, des standards. Parmi ceux-ci, deux concernent la sécurité des systèmes biométriques : le standard **19792** *Technologies de l'information — Techniques de sécurité — Cadre de la sécurité pour l'évaluation et le test de la technologie biométrique* et le standard **24745** *Technologies de l'information – Techniques de sécurité – Protection des informations biométriques*. Le premier couvre les aspects spécifiques à la biométrie et les principes qui doivent être pris en compte lors de l'évaluation de la sécurité d'un système biométrique. Le second détaille : les exigences de sécurité (confidentialité, intégrité, disponibilité, renouvellement et révocabilité) ; les exigences de respect de la vie privée (irréversibilité, non-associativité, confidentialité) ; les exigences et les lignes directrices pour la gestion et le traitement des informations biométriques garantissant la sécurisé et le respect de la vie privée.
- Le sous-comité ISO/IEC JTC1-SC37 *Biométrie* [5].
  Le périmètre du SC37 est la normalisation des technologies biométriques génériques ayant trait aux personnes en vue de prendre en charge l'interopérabilité et l'échange de données entre applications et systèmes.

Au niveau de l'Union Européenne, on trouve le CEN (Comité Européen de Normalisation), en charge d'harmoniser les normes entre les différents membres. Le Comité Technique 224 (*Technical Committee, TC*) couvre les projets de standardisation en relation avec *les applications biométriques conviviales pour les utilisateurs*. Son intitulé exact est le suivant : Identification des personnes et dispositifs à caractère personnel associés, comprenant élément de sécurité, systèmes, opérations et données privées sécurisés dans un environnement multisectoriel. L'ENISA (European Union Agency for Cybersecurity) propose que les exigences de sécurité soient sélectionnées au moyen d'une procédure d'évaluation des risques afin de : garantir que les objectifs de sécurité déterminés pour l'utilisation prévue de la cible sont atteints ; que les menaces potentielles sont correctement atténuées par les exigences de sécurité établies [ENI19].
Comment s'organisent tous ces niveaux de normalisation ? La réponse, pour la France, vient du

4. https://www.iso.org/fr/committee/45306.html
5. https://www.iso.org/fr/committee/313770.html

site de l'AFNOR (Association française de normalisation) [6] :
« Les commissions de normalisation sont animées par les bureaux de normalisation sectoriels ou par AFNOR, qui assure également la coordination d'ensemble. À l'échelle internationale, AFNOR défend les intérêts français en tant que membre des associations de normalisation européenne (CEN et CENELEC) et internationale (ISO et IEC). Son influence y est à la fois technique et stratégique, essentielle pour les entreprises françaises car 90% des normes appliquées en France sont d'origine internationale. »
En fait, les normes publiées par les organismes européens de normalisation (dont le CEN) sont « de droit » des « normes françaises homologuées » (Décret n°2009-697 du 16 juin 2009 relatif à la normalisation).
En pratique, ces contraintes réglementaires se traduisent par des analyses de risques et des cartographies de vulnérabilités, soit en considérant le système d'authentification biométrique comme un élément d'une architecture de sécurité, soit centrées sur le processus d'authentification et les données biométriques. Ces deux approches sont complémentaires et font l'objet des deux prochains paragraphes.

2.2 Analyses de risques et cartographies des vulnérabilités d'un système biométrique

Peu d'articles abordent l'étude de sécurité d'un système biométrique en termes d'analyse de risques. On peut mentionner quelques références parmi lesquelles [Brö06], [Rob07], [El-+12], [Fer13], [JMD20]. Pourtant, il semble essentiel de prendre du recul et de ne pas se contenter d'étudier la sécurité d'un système biométrique de façon isolée. Un système d'authentification (ou identification) est nécessairement au cœur d'un système d'information plus vaste. Il semble donc pertinent d'utiliser des outils standard d'analyse de risques dans une première étape. Cette approche fait le lien avec mon cours de 3A à l'ENSICAEN, dans le module Sécurité des Systèmes d'Information, sur la Cybersécurité et la gestion des risques : ce cours est centré sur la sécurité informatique (réglementations et instances européennes et françaises, définitions, critères, etc.), la notion de risque (définitions, normes ISO 2700x, gestion des risques, etc.) et les méthodes d'analyse de risques (NIST et EBIOS). C'est pourquoi on retrouve de nombreux éléments et outils en commun entre mon cours et cette partie du manuscrit.
Pour commencer, les définitions suivantes sont détaillées dans l'article [Brö06] déjà cité, adaptées de la terminologie standard du domaine de la sécurité des systèmes d'information, pour les systèmes d'authentification biométrique :

Définition 1. *Une **menace** pour le système d'authentification biométrique est la possibilité qu'un événement ou une action entraîne une perte de sécurité, une dégradation de la fiabilité ou des performances de la technologie, ou l'atteinte à la vie privée d'une personne.*

Définition 2. *Une **vulnérabilité** est une faiblesse d'un bien, qui peut être exploitée par une ou plusieurs menaces. La vulnérabilité d'un système biométrique est définie comme la possibilité d'une attaque contre un système biométrique, par un attaquant actif.*

Définition 3. *Un **risque** spécifique pour la technologie d'authentification biométrique est la probabilité qu'une menace spécifique à cette technologie soit exploitée contre une vulnérabilité spécifique elle aussi, avec des conséquences et des effets potentiellement néfastes.*

Roberts précise qu'il existe trois dimensions-clés dans les attaques, et chaque dimension requiert un traitement différent [Rob07]. Il s'agit :

- des agents menaçants

6. https://normalisation.afnor.org/

- des vecteurs de menaces ou points d'attaque
- des vulnérabilités du système

Un *agent menaçant* peut être soit un **imposteur** (une personne qui intentionnellement ou non se fait passer pour un utilisateur légitime), un **attaquant** (un système ou une personne qui tente de compromettre le système biométrique), ou un **utilisateur légitime** (maladroit). L'article de Jain *et al.* [JRN11] va plus loin et ajoute des agents menaçants parmi les caractéristiques du système biométrique lui-même, à savoir : les **limitations intrinsèques** (celles des composants du système, influant sur les taux de FMR et FNMR, vus au chapitre précédent), et les **adversaires** (qui peuvent manipuler le système grâce à un accès physique ou logique). Les auteurs déclinent les menaces pour la sécurité des système biométriques en quatre types : le déni de service, l'intrusion, le rejet et le détournement de finalité qui peut avoir des impacts sur la sécurité et le respect de la vie privée.

Si on revient à l'article de Brömme [Brö06], l'auteur adopte une approche holistique d'analyse des risques de sécurité. Il propose un modèle des systèmes d'authentification biométrique adapté à une analyse de risques, risques qui se situent à différents niveaux :

- risques de capture
- risques de transmission
- risques de stockage
- risques de calcul

En accord avec les approches standard de gestion des risques, cet article propose une matrice de risques pour les méthodes d'authentification biométrique, centrée sur les trois étapes : enrôlement, authentification, désenrôlement. L'auteur en conclut qu'une approche holistique de la sécurité permet le développement de technologies d'authentification biométrique plus sûres.

Concernant les modèles de menaces existants, on peut en trouver une présentation détaillée dans l'article récent de Joshi *et al.* [JMD20], qui propose également un nouveau modèle.
Les quatre modèles existants sont les suivants :

- **le modèle de Ratha *et al.*** [RCB01], présenté à la figure 1.1. Ce modèle fait toujours

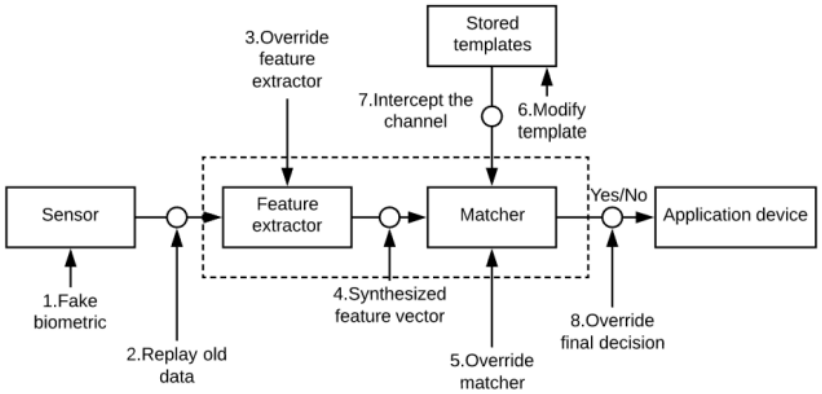


FIGURE 1.1 – Modèle de Ratha [RCB01]

référence (je le cite dans mon cours de M2 sur la protection de la biométrie). Il identifie huit points d'attaques génériques au niveau des différents modules constitutifs d'un système biométrique : le capteur, l'extracteur de caractéristiques, la base de données, le module de comparaison.

- **le modèle *fishbone*** [JNN08] présenté à la figure 1.2. Ce modèle propose de répartir les

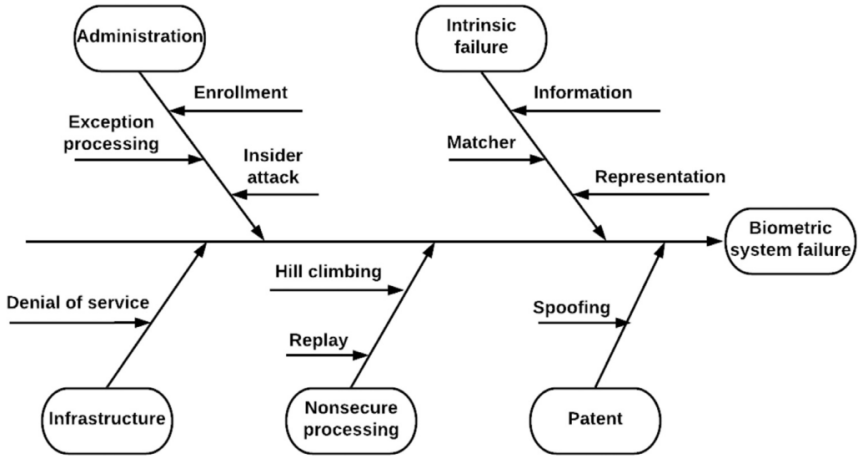


FIGURE 1.2 – Modèle fishbone [JNN08]

points d'attaques sur une autre organisation des composants d'un système biométrique : l'administration, l'infrastructure, le traitement non sécurisé des données, les faiblesses intrinsèques, etc. Il présente principalement les différentes causes menant à l'échec du système.

- **le modèle de Nagar *et al.*** [JNN08] présenté à la figure 1.3. Ce modèle, établi sur la

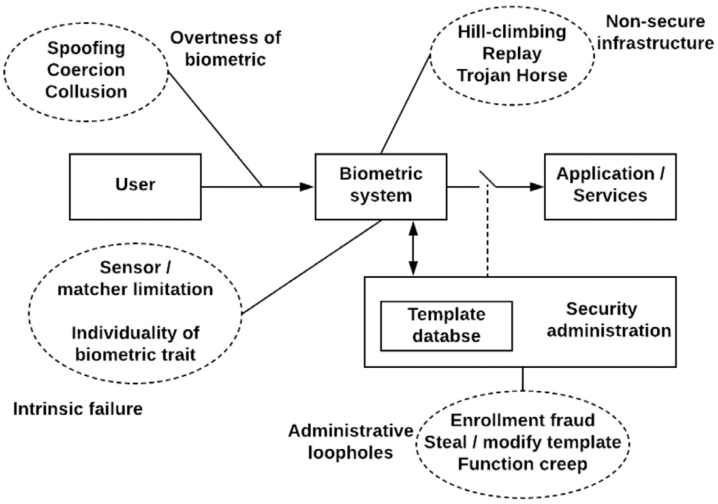


FIGURE 1.3 – Modèle de Nagar [JNN08]

base du modèle *fishbone*, regroupe les points d'attaque du modèle de Ratha en quatre catégories :

  - attaques à l'interface avec l'utilisateur
  - attaques au niveau des interactions entre les modules (ou composants)
  - attaques sur les modules
  - attaques sur la base de données des gabarits

Il replace le système biométrique au sein de l'architecture prenant en compte l'utilisateur et le service auquel il permet d'accéder.

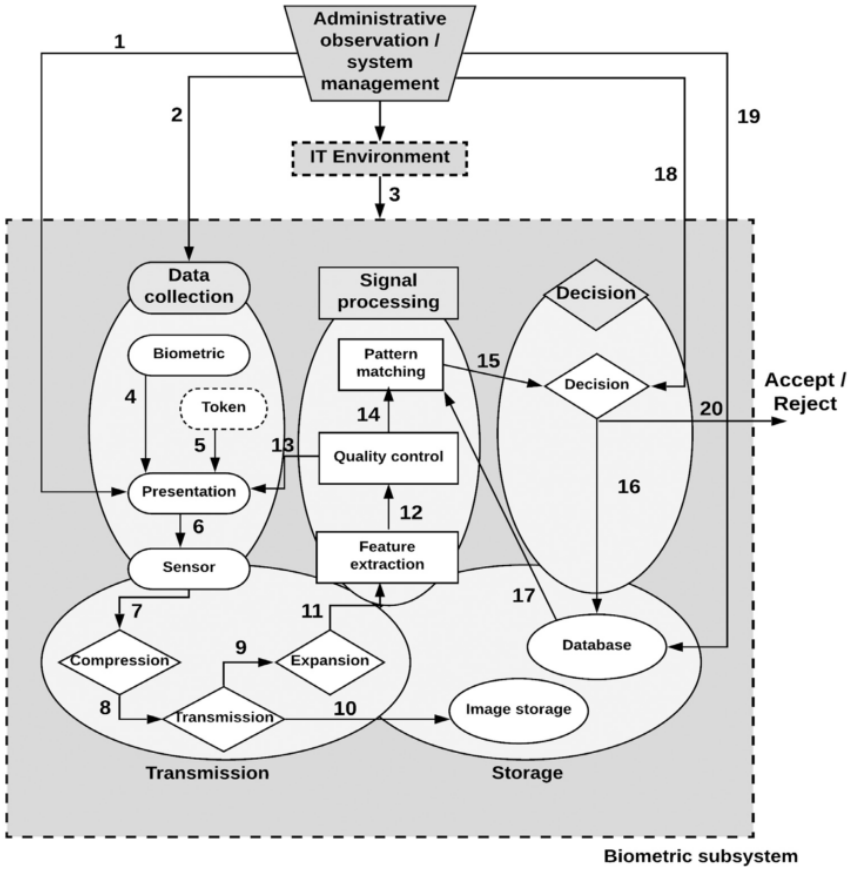


FIGURE 1.4 – Modèle de Bartlow et Cukic [BC05]

- **le modèle de Bartlow et Cukic** [BC05] présenté à la figure 1.4. La base de ce modèle est celui de Ratha, combiné à l'architecture de Wayman [Way96]. Cette architecture propose de décomposer un système biométrique en cinq sous-systèmes, en fonction de leur rôle : collecte de données, transmission, traitement du signal, stockage, décision. Le modèle de Bartlow et Cukic comporte trois grands modules (le sous-système biométrique – qui regroupe les cinq composants de Wayman –, l'environnement IT, et la gestion administrative du système). Il distingue une vingtaine de points d'attaque et vingt-deux vulnérabilités. Il est beaucoup plus détaillé que les autres modèles, et de ce fait, semble difficilement applicable en pratique.

A partir des quatre modèles existants, Joshi *et al.* ont effectué dans l'article [JMD20] une mise à jour des points d'attaque publiés dans la littérature récente pour proposer le modèle de la figure 1.5. Ce modèle est valable pour un système d'authentification (par empreinte digitale dans l'article) avec comparaison dans la base de données.

Il peut être intéressant de regarder les spécificités des systèmes biométriques et leurs vulnéra-

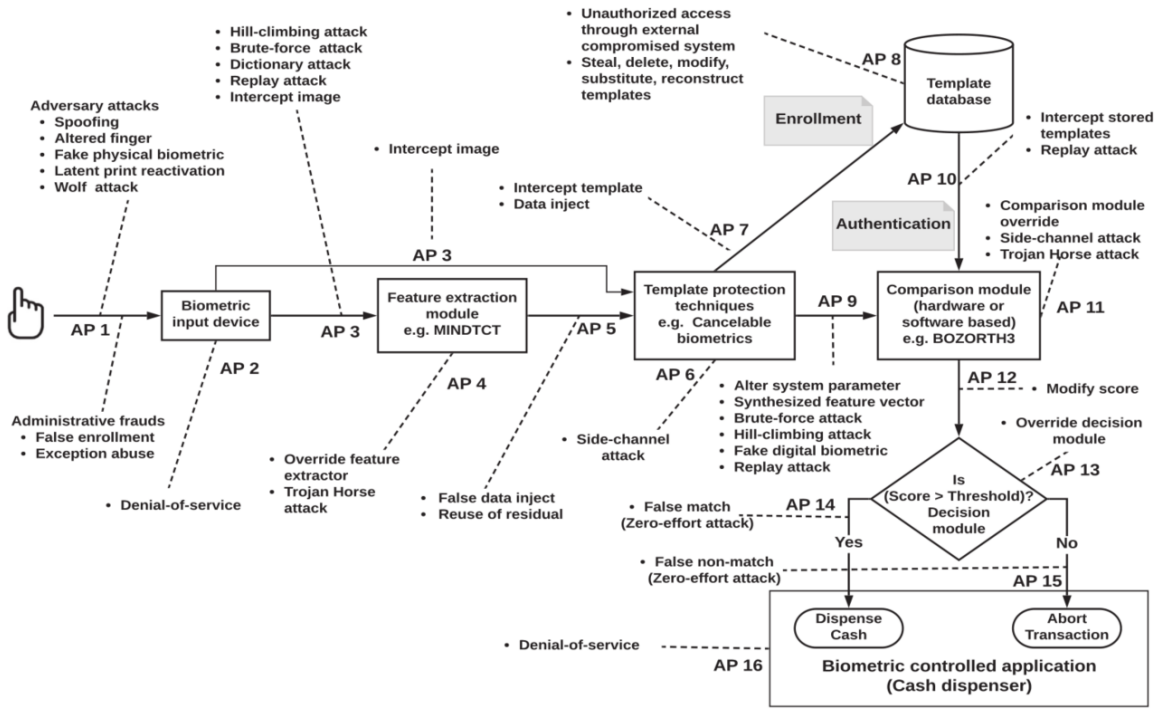


FIGURE 1.5 – Nouveau modèle proposé par Joshi *et al.* [JMD20]

bilités intrinsèques. Ainsi, on trouve une autre classification des évaluations de sécurité dans le manuscrit de thèse de M.B. Fernández Saavedra [Fer13], classification établie selon les Critères Communs :

- les tests de conformité pour évaluer si les exigences de sécurité sont satisfaites ou non ;
- l'évaluation des vulnérabilités : dresser une liste des menaces potentielles, décider lesquelles sont exploitables et concevoir des contre-mesures spécifiques

L'article de El Abed *et al.* [El-+12] propose de suivre une démarche inspirée de la méthodologie EBIOS [7] (Expression des Besoins et Identification des Objectifs de Sécurité) pour calculer un niveau de risque indépendant de la modalité biométrique considérée.

Le recours au formalisme des Critères Communs ou de l'analyse des risques EBIOS peut se révéler assez contraignante, mais permet de réfléchir de façon structurée aux sources de menaces, les impacts des événements redoutés, les probabilités d'occurrence des différentes menaces, selon les différents critères de sécurités considérés. Ces méthodes présentent l'intérêt de s'adapter à chaque système particulier, en complément des cartographies standardisées et donc figées précédentes.

Pour terminer sur ces aspects, la figure 1.6, extraite de la référence [Cam13], dresse une liste (non-exhaustive) de vulnérabilités reconnues des systèmes biométriques.

Les méthodes classiques d'analyse de risques contribuent donc à l'étude de sécurité des systèmes d'authentification biométrique, considérés en tant que système d'information, en établissant une cartographie des vulnérabilités. Cette cartographie dépend du modèle retenu et constitue un guide dans la conception de mesures préventives ou protectrices, diminuant ainsi la probabilité d'occurrence ou la gravité des risques associés. Il ne faut toutefois pas oublier que les données biométriques sont des données particulières, sensibles, personnelles (*cf.* le chapitre 3) dont les

7. https://www.ssi.gouv.fr/guide/ebios-2010-expression-des-besoins-et-identification-des-objectifs-de-securite/

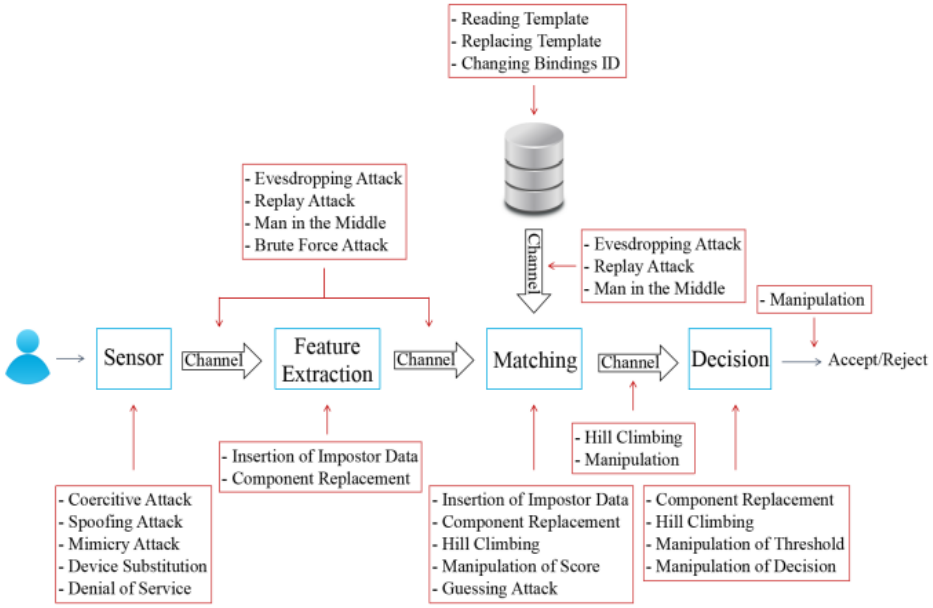


FIGURE 1.6 – Différents types d'attaques [Cam13]

besoins de sécurité sont maximaux.

2.3 Sécurité des données biométriques

Dans ce paragraphe, sont concernés les risques relatifs aux données biométriques : par nature, ce sont des données personnelles, sensibles, non révocables. Des risques spécifiques émergent du besoin impératif de protection des données biométriques : risques concernant le stockage ou la comparaison des données ; risques liés à l'usage du système d'authentification biométrique ; risques en lien avec la protection des données [JNN08]. Ce dernier point sera au cœur du chapitre 3 et ne sera donc pas abordé ici.
Les processus de stockage et de comparaison dépendent de l'architecture du système biométrique étudié. En général, l'architecture détermine comment les différents composants d'un système s'organisent et s'intègrent, décrit où se situent les flux et les transferts de données, où elles sont stockées et/ou accessibles. Au sein de l'architecture d'un système biométrique, les données biométriques capturées transitent par différents processus : capture, traitement, extraction des caractéristiques, stockage, comparaison. Tout au long de cette chaîne de traitement, des erreurs peuvent survenir : un utilisateur légitime peut être rejeté à tort (générant par exemple un risque sur sa liberté de mouvement), un imposteur peut être accepté à tort (induisant par exemple un risque de sécurité en matière de disponibilité, intégrité, une usurpation d'identité, un vol de données, etc.) Par ailleurs, on constate que des vulnérabilités sont liées à la nature même des systèmes biométriques : ce ne sont pas des systèmes déterministes, dans le sens où la donnée présentée n'est jamais exactement la même que celle stockée comme référence. Les systèmes biométriques reposent sur une comparaison, puis une décision, donc une prise de risque : les attaquants peuvent exploiter cette marge de risque qui existe toujours, aussi réduite soit-elle. De bonnes performances (en termes de taux de reconnaissance ou d'EER) sont donc nécessaires, mais pas suffisantes à garantir la sécurité des données.
Il faut souligner qu'une distinction importante existe entre des systèmes qui stockent les données dans une base centralisée et d'autres systèmes qui stockent les données dans un objet personnel

qui reste en permanence sous le contrôle de son propriétaire. La base de données centralisée comporte plusieurs vulnérabilités (*cf.* la figure 1.6) : la référence d'un utilisateur légitime peut être remplacée par celle d'un imposteur ; une donnée biométrique peut être contrefaite à partir d'une référence légitime ; la référence compromise peut être rejouée auprès d'un autre système, pour usurper l'identité de l'utilisateur de départ. Comme nous le verrons au chapitre 3, en raison du caractère non révocable des données biométriques, de leur lien unique avec l'utilisateur, de leur universalité, le vol ou la compromission d'une telle donnée peut affecter gravement et définitivement la vie privée des utilisateurs enrôlés dans le système.
Se pose alors la question de la protection des références (ou gabarits) stockés à des fins d'authentification ultérieure : en effet, si les données biométriques brutes constituent sans nul doute des données personnelles sensibles, qu'en est-il des caractéristiques extraites de données qui ont éventuellement subi un pré-traitement (normalisation, ré-échantillonnage) ? La question est débattue dans la thèse d'Els Kindt [Kin12], qui s'inquiète du risque que « *toute personne soit toujours enrôlée quelque part* », la raison étant le nombre croissant de bases de données privées. Elle souligne également que, si la standardisation (réalisée par les normes internationales, voir le paragraphe 2.1) améliore l'interopérabilité des systèmes, dans le même temps, elle rend l'authentification ou l'identification des citoyens plus aisée par les autorités publiques (police, justice), même sur la base de gabarits. D'où le besoin de protection de ces modèles de référence, au même titre que les données biométriques brutes.

2.4 Discussion

La première partie de ce chapitre a permis de démêler les liens historiques et complexes entre les systèmes d'authentification biométrique et la sécurité. L'intrication des deux domaines a constamment évolué au cours de l'histoire, cette évolution s'accélérant avec le développement des nouvelles technologies depuis une vingtaine d'années. Les données biométriques constituent des données personnelles sensibles, et à ce titre ont des besoins de sécurité élevés. Ces besoins se traduisent par des normes, des analyses de risques, des cartographies de vulnérabilités, de points d'attaques, plus ou moins génériques. Les études de sécurité doivent également se concentrer sur le système biométrique et sur ses besoins particuliers, au niveau de son architecture, ses performances, ou encore la protection des données concernées. La partie suivante relate deux contributions principales tirées des thèses de Syed Zulkarnain Syed Idrus [Sye14] et d'Abir Mhenni [Mhe19], en lien avec ces thématiques.

3 Contributions

On vient de voir que l'évaluation de la sécurité d'un système biométrique repose notamment sur celle de ses performances : dans l'idéal, le système laisse passer systématiquement les utilisateurs légitimes, et rejette tous les imposteurs. Ce cas décrit un système ayant un EER (Equal Error Rate) égal à zéro. L'EER est en pratique une mesure standardisée de la performance d'un système biométrique, qui correspond à un choix du seuil de décision où la condition suivante est vérifiée : False Match Rate = False Non Match Rate. Par conséquent, plus l'EER est faible, meilleures sont les performances du système biométrique (voir le chapitre d'introduction page 1).

Même si l'EER est rarement égal à zéro, on a vu que la biométrie possède, par rapport aux autres moyens d'authentification, un avantage certain : elle permet d'authentifier l'utilisateur lui-même, et non un objet qu'il possède, ni une information, un élément secret qu'il connaît. Un autre avantage, non négligeable, se situe du côté de l'utilisateur, qui n'a rien à retenir, ni à

apporter pour utiliser un système biométrique. L'article [Sye+13] présente un tour d'horizon des différentes méthodes d'authentification.
Différentes approches sont présentes dans la littérature, qui contribuent à une amélioration des performances des systèmes biométriques :

- l'amélioration du matériel, des capteurs, des conditions de capture des données biométriques
- l'évaluation de la qualité des données biométriques
- la multibiométrie
- la biométrie douce
- la mise à jour de modèle

Les travaux réalisés dans le cadre des thèses que j'ai encadrées m'ont amenée à étudier les deux dernières pistes, que je vais détailler dans la suite de cette partie. En dehors de ces thèses, j'ai participé à la rédaction d'un chapitre de livre consacré à la multibiométrie [Gio+12].

3.1 La biométrie douce

La notion de biométrie douce a été introduite par Jain *et al.* en 2004. Elle est définie de la manière suivante : *un caractère fournissant de l'information sur l'individu, mais manquant d'unicité et de permanence pour différencier suffisamment deux individus*. Ce sont donc des caractéristiques, appelée *traits*, qui ne sont pas suffisantes pour authentifier un individu, mais peuvent aider à la construction d'un profil, comme par exemple, à partir de la capture du visage d'un utilisateur : la couleur de sa peau, la couleur de ses cheveux, la couleur de ses yeux, etc. Plus généralement, la biométrie douce repose sur des informations extraites d'une modalité biométrique, comme le sexe, l'âge, la taille, le poids, la démarche, ou d'autres mesures du corps. On renoue ici avec l'anthropométrie telle que définie par Bertillon, aux origines des systèmes d'authentification modernes. La biométrie douce peut être exploitée de façon isolée, ou combinée avec un résultat d'authentification biométrique classique.
Grâce à l'extraction de caractéristiques supplémentaires, un système biométrique combiné à de la biométrie douce présente naturellement de meilleures performances et pourrait être considéré comme un système multibiométrique : la multibiométrie regroupe un ensemble de techniques, ayant pour point commun l'utilisation de plusieurs captures de données biométriques, chaque capture s'effectuant dans un sous-module. Un algorithme de fusion des résultats de chaque sous-module peut être appliqué à différents niveaux : fusion de caractéristiques, fusion de scores, de rangs, de décisions. Automatiquement, de meilleures performances augmentent la confiance dans le résultat produit par le système d'authentification. La sécurité est renforcée lorsque l'utilisateur doit présenter plusieurs données biométriques : en effet, l'usurpation d'identité est rendue compliquée pour un attaquant, s'il doit présenter des données falsifiées à un système de reconnaissance d'iris puis un système à base d'empreinte digitale, ou à un système de reconnaissance d'empreinte digitale à base d'empreintes de plusieurs doigts, etc... Les systèmes multibiométriques représentent donc une parade assez naturelle contre les attaques de présentation. Tous ces aspects sont déjà présents dans le chapitre [Gio+12], publié en 2012, ou dans les articles plus récents [DH17], [SSR19], qui proposent un tour d'horizon et une mise en perspective des applications de la multibiométrie. Il faut noter toutefois que l'utilisation successive de plusieurs capteurs biométriques représente une dégradation de l'expérience utilisateur.
Il existe une autre différence majeure entre la multibiométrie et la biométrie douce : chaque trait de biométrie douce se décline en plusieurs instances (par exemple la couleur des cheveux, du blond très clair, au noir, en passant par le roux, le châtain, le brun, le gris, ou l'absence de

cheveux) et la frontière entre ces instances peut être délicate à déterminer dans certains cas ; en revanche, pour un système multibiométrique, chaque sous-système est dédié à une modalité biométrique standard, avec un module de décision dédié. Le travail de conception, de décision et d'évaluation des performances est donc fondamentalement différent dans les deux cas. En outre, un système de biométrie douce n'inclut pas nécessairement de phase d'enrôlement : le système peut apprendre à reconnaître des cheveux roux sans avoir enrôlé un nouvel utilisateur. Cependant ce n'est pas le cas pour tous les traits de biométrie douce, comme on va le voir dans la thèse de Syed Zulkarnain Syed Idrus.

Syed Zulkarnain a obtenu une bourse de la Malaisie en 2011 pour une thèse, co-dirigée par Christophe Rosenberger et Patrick Bours, que j'ai co-encadrée. Le titre de la thèse est *Soft Biometrics for Keystroke Dynamics*, elle a été soutenue en 2014. La dynamique de frappe au clavier (DDF), ou frappologie, est une modalité comportementale étudiée depuis les années 1980, qui cependant ne trouve son essor qu'à partir des années 2000. Elle possède un certain nombre d'atouts : pas de coût supplémentaire et une sécurité renforcée (par rapport à un mot de passe seul), pas de contrainte supplémentaire pour l'utilisateur. Les techniques de biométrie douce sont peu appliquées aux modalités comportementales, qui souffrent pourtant de performances de reconnaissance plus faible que les modalités classiques, physiques (visage, empreinte digitale, iris, etc.). Lorsque Syed Zulkarnain commence sa thèse en 2011, il y a très peu de littérature sur la biométrie douce pour la DDF. On peut mentionner les travaux de Epp *et al.* [ELM11], sur la reconnaissance de l'état d'esprit de l'utilisateur en fonction de sa façon de taper au clavier. Dans la thèse de Romain Giot [Gio12], effectuée au sein du laboratoire GREYC, est étudiée la reconnaissance du sexe de l'utilisateur, toujours en fonction de la DDF. La thèse de Syed Zulkarnain mentionne quelques références supplémentaires (pages 71-72).

Dans cette thèse, les traits de biométrie douce retenus sont en lien avec l'action de taper sur un clavier d'ordinateur :

- **T1** - l'utilisateur est droitier/gaucher ;
- **T2** - l'utilisateur tape au clavier avec une/deux main(s) ;
- **T3** - l'utilisateur est un homme/une femme ;
- **T4** - l'utilisateur a moins/plus de trente ans.

L'extraction de traits de biométrie douce se fait au niveau logiciel uniquement, en parallèle du processus d'authentification biométrique par DDF. Il n'y a donc pas de dégradation de l'expérience utilisateur en lien avec la biométrie douce.

La dynamique de frappe au clavier repose sur la collecte d'événements qui ont lieu lors de la frappe sur un clavier d'ordinateur : généralement la pression et le relâchement de touches successives. Y sont associés des temps (de pression, de relâchement, de latence, de vol), comme illustré à la figure 1.7.

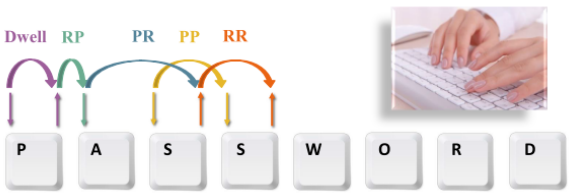


FIGURE 1.7 – Caractéristiques de la dynamique de frappe au clavier

Les données extraites sont les suivantes :

- le temps de pression des touches **PR**
  Il se calcule comme la différence entre l'instant du relâchement de la touche et l'instant de sa pression
- le temps de latence entre l'appui des touches **PP**
  Il se calcule comme la différence entre l'instant de la pression d'une touche et l'instant de la pression de la touche précédente
- le temps de vol **RP**
  Il se calcule comme la différence de temps entre le relâchement d'une touche et la pression d'une autre
- le temps entre deux relâchements de touches **RR**

Pour un mot de passe statique, la référence (le gabarit) sera construite à partir de ces différents temps. On peut appliquer auparavant des pré-traitements (normalisation, suppression de valeurs aberrantes liées à une hésitation, quantification des valeurs possibles). Il est également possible de travailler sur des données issues de texte libre : dans ce cas, on étudie les **digraphes**, plus précisément les temps de latences entre deux pressions successives (ou PP, *cf.* la figure 1.7). Ces deux types d'authentification sont complémentaires :

- l'authentification par DDF à base de mot de passe fixe apporte un niveau de sécurité supplémentaire par rapport à l'authentification standard par mot de passe ;
- l'authentification par DDF à base de texte libre permet une authentification continue de l'utilisateur, transparente, et ce, tout au long de sa session de travail sur ordinateur.

Les performances de la DDF se situent classiquement aux environs de 10%, elles sont donc nettement inférieures à celles des modalités physique ou physiologiques. La biométrie douce, au cœur de la thèse de Syed Zulkarnain, paraissait être une piste intéressante pour l'amélioration de ces performances. Étant donné le contexte favorable de la co-direction franco-norvégienne de sa thèse, il a collecté des données auprès de 110 volontaires français et norvégiens. Le tableau 1.1, extrait de la thèse de Syed Zulkarnain [Sye14], décrit les caractéristiques des données de cette base conséquente, mise à disposition de la communauté scientifique.

La constitution de cette base représente la première contribution de la thèse de Syed Zulkarnain : en effet, ce type de données est assez difficile à obtenir (soit à collecter, soit à télécharger en tant que base de données publique). C'est pourquoi la collecte a été réalisée dès le début de la thèse [Sye+14]. En raison des origines géographiques variées des utilisateurs – elles sont le reflet de la diversité des membres des laboratoires de recherche –, et par conséquent des différences culturelles, le scénario d'acquisition retenu se compose de cinq mots de passe connus par tous : *leonardo dicaprio*, *the rolling stones*, *michael schumacher*, *red hot chilli peppers*, *united states of america* (en minuscules).

Parmi les contributions de la thèse de Syed Zulkarnain, je retiens les trois suivantes, en lien direct avec la biométrie douce :

- il a suivi deux pistes pour évaluer la reconnaissance de traits de biométrie douce pour la DDF :
  - la première repose sur les mots de passe définis juste avant ;
  - la seconde exploite les mêmes mots de passe sous forme de texte libre ;
- il a amélioré les performances du système d'authentification par DDF seule en prenant en compte des traits de biométrie douce

Je présente ici uniquement les contributions sur la reconnaissance de traits de biométrie douce, à partir de texte fixe, ou de texte libre.

Information	Description
Number of users	110
Users from France	70
Users from Norway	40
Users' country of origin	France, Norway, Netherlands, Germany, Denmark, Spain, Greece, Ukraine, Iran, Czech Republic, Serbia, Syria, Lithuania, Bulgaria, Mali, Lebanon, India, Vietnam, Malaysia, Indonesia, China, Japan, New Zealand and United States of America.
Gender	78 males (47 from France, 31 from Norway); and 32 females (23 from France, 9 from Norway)
Age range	Between 15 and 65 years old
Age classes	$<$ 30 years old (37 males, 14 females); and $\geq$ 30 years old (41 males, 18 females)
Handedness	98 right-handed (70 males, 28 females); and 12 left-handed (8 males, 4 females)
Number of known passwords	5
Database sample length	17 characters ("leonardo dicaprio") 18 characters ("the rolling stones") 18 characters ("michael schumacher") 22 characters ("red hot chilli peppers") 24 characters ("united states of america")
Database sample size	11,000 data (= 5 passwords x 2 classes x 110 users x 10 entries)
Typing error	Not allowed
Controlled acquisition	Yes
User profession	Students, researchers, faculty members, administration staff, others (housewifes/non-working people)
Keyboard	2 external keyboards: AZERTY & QWERTY
Acquisition platform	Windows XP & GREYC keystroke software

Tableau 1.1 – Description des données collectées en France et en Norvège [Sye14]

Préparation des données

Chaque volontaire a tapé chacun des 5 mots de passe, 10 fois avec deux mains (conjointement) et 10 fois avec une seule main. Les données collectées en tapant avec une seule main ne sont utilisées que pour la reconnaissance du trait *T2 - l'utilisateur tape au clavier avec une/deux main(s)*. Pour toutes les autres études, on utilise uniquement la frappe au clavier avec deux mains, qui correspond au comportement naturel des volontaires. Pour chaque utilisateur, les trois premières captures sont écartées pour éviter « l'effet découverte » de chaque mot de passe. Les données restantes sont ainsi plus consistantes.

Parmi les traits de biométrie douce retenus, certains induisent des classes déséquilibrées : il y a plus d'hommes (78) que de femmes (32) et il y a plus de droitiers (98) que de gauchers

(12). On a choisi de rééquilibrer les classes par un sous-échantillonnage aléatoire de la classe majoritaire. On a utilisé la technique de Bootstrapping : l'opération a été répétée une centaine de fois, et les résultats présentés sont la moyenne de ces cent itérations.
La répartition des données entre jeu d'apprentissage et jeu de test a également été analysée : pour chaque scénario retenu, Syed Zulkarnain a étudié le taux de reconnaissance d'un trait de biométrie douce pour un pourcentage de données d'apprentissage fixé entre 1% et 90%. On obtient alors des courbes représentant le taux de reconnaissance exacte en fonction du pourcentage de données réservées à l'apprentissage, dont l'allure est conforme à ce type de courbe, comme illustré à la figure 1.8.

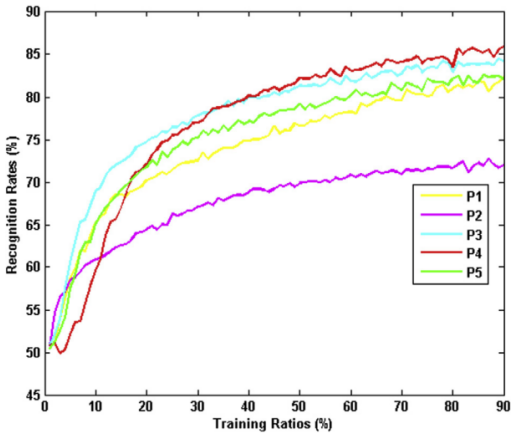


FIGURE 1.8 – Taux de reconnaissance du trait « T3 - l'utilisateur est un homme/une femme »

Biométrie douce, mots de passe statiques et texte libre

Dans la thèse de Syed Zulkarnain Syed Idrus, les données collectées à partir des cinq mots de passe ont été exploitées pour :

- la reconnaissance des traits de biométrie douce T1 à T4 à partir des mots de passe considérés comme fixes, avec l'entraînement d'un SVM par trait ;
- l'amélioration des performances par fusion des scores ou des décisions de chaque SVM ;
- la reconnaissance des traits de biométrie douce T1 à T4 à partir des digraphes extraits des mots de passe, considérés comme du texte libre : on ne conserve que les digraphes ayant une occurrence supérieure ou égale à deux, parmi tous les digraphes présents dans les cinq mots de passe.

Les résultats obtenus pour le texte fixe sont reportés dans le tableau 1.2. Dans ces différents scénarios, on a conservé 50% des données pour l'apprentissage, pour chaque SVM.
Les résultats obtenus pour le texte libre sont reportés dans le tableau 1.3. L'intervalle correspond aux performances réalisées avec une proportion de données d'apprentissage entre 50 et 90% des données totales.
On constate que le trait *T4 - l'utilisateur a moins/plus de trente ans* est le plus difficile à reconnaître. Tous les autres affichent une reconnaissance supérieure à 80% (avec fusion des cinq mots de passe dans le cas du texte fixe). La cause en est peut-être le faible impact de l'âge sur la façon de taper : le rapport à l'ordinateur, sa durée d'utilisation quotidienne influence sans doute plus sûrement la DDF d'un utilisateur. Ces informations n'ont pas été collectées au moment de la capture des données. Néanmoins, s'agissant d'une collecte au sein de deux laboratoires de

		By fusing	
Soft category	**Without fusion**	**Majority voting**	**Score fusion**
Hand	94%	100%	100%
Gender	63%	86%	92%
Age	55%	87%	86%
Handedness	62%	85%	92%

Tableau 1.2 – Reconnaissance des traits de biométrie douce sans et avec fusion (extrait de [Sye14])

Soft category	**Free text (in %)**
Hand	[97,98]
Gender	[79,84]
Age	[72,75]
Handedness	[83,88]

Tableau 1.3 – Reconnaissance des traits de biométrie douce avec texte libre (extrait de [Sye14])

recherche en information, on peut imaginer que tous les volontaires sont habitués à l'outil informatique.

Les résultats de création d'un profil des utilisateurs obtenus par Syed Zulkarnain à partir de la collecte d'une faible quantité de données de DDF laissent deviner l'ampleur des prédictions sur nos comportements par les GAFAM et autres BATX à partir de nos traces numériques abondantes et variées. Ces géants du net récupèrent bien plus que les informations relatives à la frappe de cinq mots de passe. Après des années à créer des profils d'utilisateurs, aujourd'hui leur défi est de prédire nos comportements à partir de l'analyse des mêmes traces, de vendre ces prédictions à des acteurs du marketing, à des agences plus ou moins opaques, etc., qui tenteront ensuite d'influencer nos comportements : la boucle est bouclée... On comprend donc l'importance de protéger les données personnelles, particulièrement les données biométriques, même des données comportementales, qui ne semblent pas très intrusives, comme la DDF : comme on l'a fait dans la thèse de Syed Zulkarnain, on pourrait sans doute essayer d'extraire d'autres traits de biométrie douce des données collectées. Le chapitre 3 propose des pistes de solutions. On peut également faire un lien avec l'utilisabilité, traitée dans le chapitre 2 : la transparence sur tout ce qui pourrait être fait avec ses données tend à renforcer la confiance de l'utilisateur dans le système biométrique. Des aspects pédagogiques, une sensibilisation, voire une éducation des utilisateurs permettrait sans nul doute de leur (re)donner une souveraineté sur leurs traces numériques. Tous ces aspects participent à l'*UX design* (conception centrée sur l'expérience utilisateur).

Perspectives

Nous ne sommes pas allés au-delà de ce que j'ai présenté ci-dessus dans la préparation ou le nettoyage des données. Les méthodes de sur-échantillonnage de la classe minoritaire, de type Smote, ou des méthodes hybrides représentent des pistes intéressantes à explorer pour contrebalancer le déséquilibre de certaines classes de données [San+17]. Il en est de même au niveau algorithmique avec l'apprentissage sensible aux coûts [VCC99], [Tin02].

3.2 La mise à jour ou adaptation de modèle biométrique

Les modalités de biométrie comportementale (dynamique de frappe au clavier, signature, démarche, schéma pour débloquer un smartphone, etc.) sont – par nature – dynamiques, et donc évoluent dans le temps. La variabilité *intra* peut ainsi se trouver affectée à l'échelle d'une journée (à cause de la fatigue, d'un stress subit), ou au fil des jours, des semaines (la façon de taper au clavier évolue constamment, l'utilisateur s'habitue à tracer son schéma de déblocage). Les points précédemment abordés peuvent fournir une solution : combiner une modalité comportementale avec une modalité physiologique (donc statique) pour créer un système multibiométrique ; renforcer l'authentification par la vérification complémentaire de traits de biométrie douce. Un ensemble de techniques paraît néanmoins conçu pour pallier ces variabilités *intra* pour le moins indésirables, techniques regroupées sous le terme de *mise à jour du modèle biométrique de référence*, ou méthodes *adaptatives*. Il s'agit de compenser la dérive temporelle inéluctable entre la référence de l'utilisateur, stockée comme modèle, et les données fraîchement présentées au système. Dans ce but, le modèle va être adapté au cours du temps : le principe initial consistait à définir un second seuil d'adaptation, plus contraignant que le seuil de décision, pour adapter le modèle, soit par remplacement pur et simple de la référence, soit par ajout de la nouvelle capture pour constituer une galerie, et ce, uniquement lorsque le système est sûr que l'utilisateur est légitime. Cette méthode, quoique simple et intuitive, présente l'inconvénient de ne conserver que des références très proches les unes des autres, elle est ainsi est incapable de prendre en compte toute l'étendue de la variabilité *intra* caractéristique de l'utilisateur : c'est l'exact contraire du but recherché.
Les travaux de la thèse d'Abir Mhenni ont porté sur l'amélioration des techniques de mise à jour de modèle biométrique. Abir a obtenu un financement de la Tunisie pour une thèse, co-dirigée par Najoua Essoukri Ben Amara et Christophe Rosenberger, que j'ai co-encadrée. Le titre de la thèse est *Contribution to the biometric template update : Application to keystroke dynamics modality*, elle a été soutenue en 2019.

Stratégies de mise à jour

Plusieurs étapes permettent de définir une véritable stratégie de mise à jour de modèle biométrique : la définition du modèle de référence ; le critère d'adaptation ; le mode d'adaptation ; la périodicité d'adaptation ; le mécanisme d'adaptation ; l'évaluation. Le manuscrit de thèse d'Abir Mhenni [Mhe19] comporte un chapitre détaillant les différentes stratégies présentes dans la littérature. On retrouve un tour d'horizon très fourni sur ce sujet également dans l'article [Pis+19]. La figure 1.9 résume ces différentes étapes. Je vais brièvement les présenter, avant la synthèse des contributions de la thèse d'Abir.
Définition du modèle de référence.
Pour capturer un maximum de variabilité *intra*, la référence ne peut pas être constituée d'une seule donnée – d'aussi bonne qualité soit-elle – au fil du temps. Cette remarque vaut pour les modalités comportementales surtout. Par conséquent, la plupart du temps, la référence est une

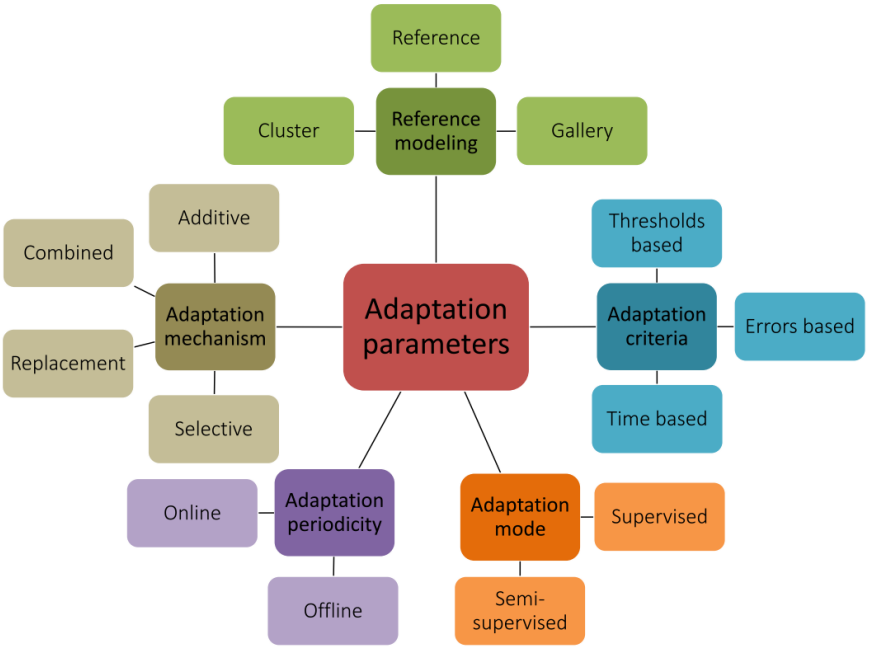


FIGURE 1.9 – Paramètres de la stratégie d'adaptation [Mhe19]

galerie constituée de plusieurs données/échantillons. Il existe plusieurs tendances pour optimiser les performances et la taille de cette galerie (*cf.* la thèse d'Abir, page 62).

Critères d'adaptation.
Le critère retenu permet de décider si le modèle de référence est mis à jour ou non. On peut citer le recours à un oracle, le processus de double seuillage, le recours à un indice de qualité, la prise en compte de la distribution temporelle des erreurs (si beaucoup de rejets se suivent, c'est peut-être la signature d'un imposteur essayant de pénétrer le système), etc.

Modes d'adaptation.
Les premiers travaux sur l'adaptation de modèle font état de mode supervisé, correspondant à une situation où les données présentées au système d'authentification sont étiquetées. Ce cas n'étant pas vraiment réaliste, seul le mode semi-supervisé est considéré aujourd'hui : il s'agit pour le système d'attribuer une étiquette (légitime ou imposteur) à la donnée fraîchement capturée. Cela se fait par auto-apprentissage (ou *self-training*, avec un seul classifieur et une seule modalité) ou par co-apprentissage (ou *co-training*, avec plusieurs modalités, souvent deux – une modalité physique ou physiologique et une modalité comportementale –, sachant que les performances de l'une sont exploitées pour l'adaptation de l'autre).

Périodicité d'adaptation.
Il existe deux cas de figure : l'adaptation différée, ou hors-ligne, et l'adaptation en temps réel, ou en-ligne. Dans les deux cas, il est possible de choisir une périodicité d'adaptation, souvent après un délai fixe, ou à chaque utilisation du système.

Mécanisme d'adaptation.
Quatre mécanismes d'adaptation existent dans la littérature :

- les mécanismes additifs : chaque nouvelle capture validée est ajoutée à la galerie (la

taille de celle-ci varie entre cinq et deux cents échantillons dans la littérature) ;
- les mécanismes de remplacement : chaque nouvelle capture validée est ajoutée à la galerie tandis qu'une ancienne est supprimée, selon un critère défini *a priori* ;
- les mécanismes multi-galeries : chaque galerie correspond à une stratégie d'adaptation différente ;
- les mécanismes de sélection : pour éviter l'explosion de la taille de la galerie, seuls les échantillons les plus *importants* (dans un certain sens) sont conservés.

Evaluation de la stratégie de mise à jour.
Le point clé à prendre en considération dans l'évaluation de la stratégie de mise à jour est la présence effective de données d'imposteurs dans les galeries des utilisateurs légitimes. Non seulement la quantité de données d'imposteurs lors de la phase d'apprentissage a un impact sur les performances futures de la mise à jour, mais également l'ordre dans lequel ces données sont présentées au système. Il est classique d'introduire environ 30% de données d'imposteurs dans le système étudié. Cependant, les attaques de type *poisoning attacks* sont rarement prises en compte : il s'agit de faire dévier le système d'une bonne reconnaissance d'un utilisateur légitime vers une amélioration de la reconnaissance d'un imposteur particulier, grâce à la présentation répétée de ses données. Un autre critère d'évaluation prend en compte la constitution des jeux de données pour l'adaptation (jeux séparés ou conjoints) lors de l'apprentissage et les tests.

Les contributions de la thèse d'Abir Mhenni que je vais présenter portent sur l'amélioration de la mise à jour du modèle pour la dynamique de frappe au clavier (DDF). Elle a, d'une part, travaillé sur l'utilisabilité, qui sera abordée dans le chapitre suivant. L'approche développée propose un enrôlement à partir d'une seule capture et un processus original d'adaptation du modèle. D'autre part, Abir a considérablement développé la personnalisation de la mise à jour du modèle, à partir de la théorie du zoo de Doddington : la stratégie d'adaptation est différente en fonction du profil de l'utilisateur, profil en lien avec la ménagerie de Doddington.

Enrôlement unique pour la mise à jour de la dynamique de frappe au clavier

L'originalité de l'approche se situe au niveau de l'enrôlement : il se fait à partir d'une seule donnée de DDF au départ, contrairement aux méthodes de l'état de l'art, qui imposent à chaque utilisateur de taper plusieurs fois son mot de passe (jusqu'à deux cents fois...). Ensuite, à chaque vérification positive, on ajoute la nouvelle donnée de l'utilisateur, jusqu'à ce que la galerie comporte dix échantillons (mécanisme additif). Ensuite un mécanisme de remplacement est mis en place pour conserver une galerie stable constituée de dix échantillons.
La figure 1.10 présente toute la stratégie détaillée par Abir dans l'article [Mhe+19b]. Le système d'authentification et de mise à jour du modèle comporte quatre blocs.

Bloc Pré-traitement.
Le template d'un utilisateur est un vecteur C, contenant les informations temporelles de la dynamique de frappe au clavier, voir la figure 1.7 : $C = [PP|PR\ RR\ RP]$. Deux traitements sont appliqués dès que la galerie comporte au moins deux échantillons. Le premier est un lissage des données aberrantes, selon la formule (1.1) : chaque composante $C(i)$ du vecteur C est comparée à trois fois la valeur de l'écart-type, calculé à partir de tous les vecteurs contenus dans la galerie. $\mu_C(i)$ est la moyenne calculée sur la i$^{\text{ème}}$composante, et $\sigma_C(i)$ l'écart-type correspondant.

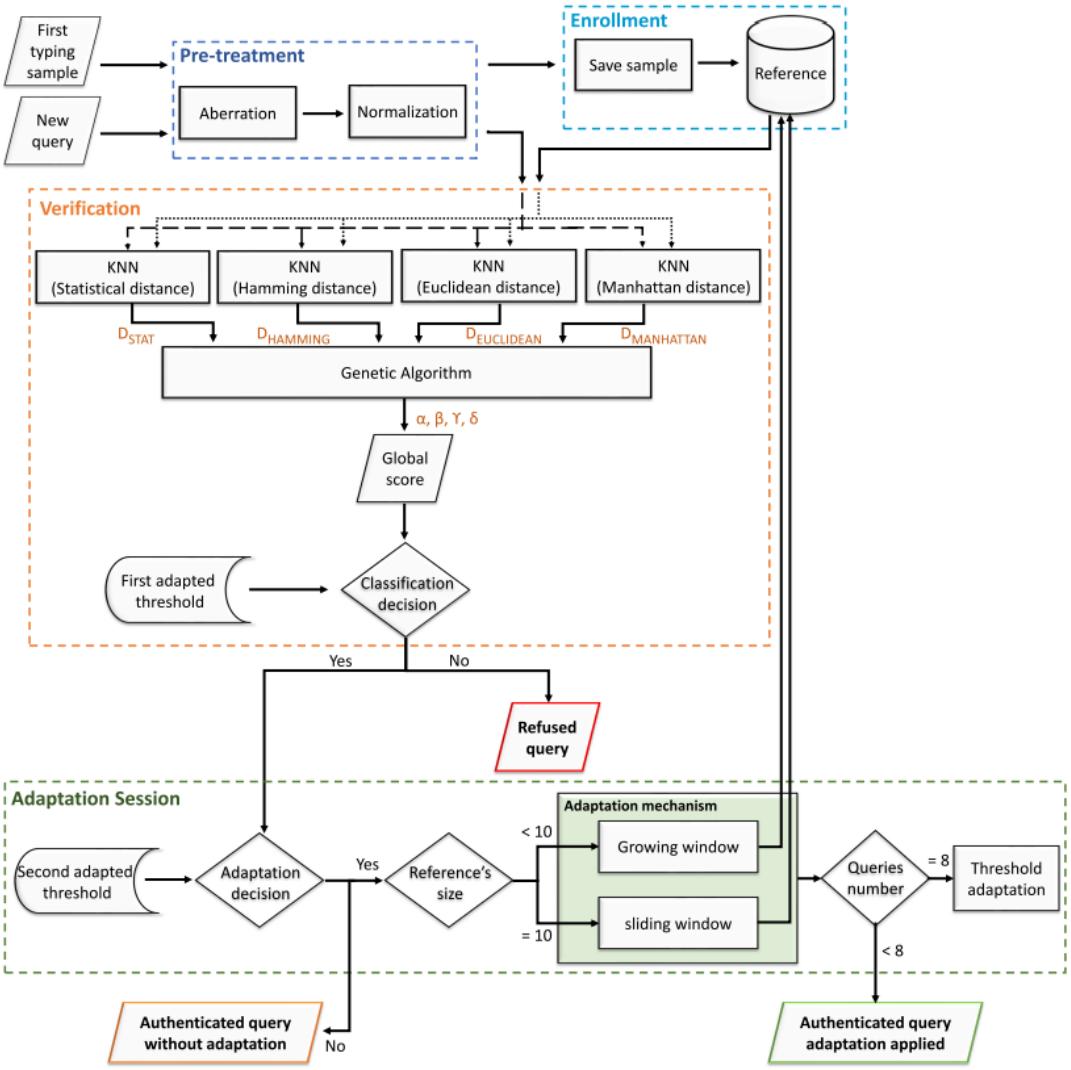


FIGURE 1.10 – Mise à jour du modèle avec un enrôlement unique [Mhe19]

$$C(i) \geq 3\sigma_C(i) \;\Rightarrow\; C(i) \leftarrow \mu_C(i) \tag{1.1}$$

Les valeurs aberrantes sont principalement dues à des hésitations, ou des perturbations subies par l'utilisateur lors de la frappe au clavier. Le second traitement est une étape de normalisation :

$$C(i) \leftarrow \frac{C(i)}{\sigma_C(i)} \tag{1.2}$$

Bloc Enrôlement.
En général, les systèmes d'authentification biométrique à base de DDF (comme pour la plupart des modalités comportementales) comportent une phase d'enrôlement composée de plusieurs captures, plus précisément, l'utilisateur doit taper plusieurs fois le mot de passe choisi. Cette habitude a pour but de compenser la variabilité intra-utilisateur. Autrement dit, cette habitude a pour but de capturer un maximum de diversité dans le comportement de l'utilisateur. Dans la plupart des articles de l'état de l'art (*cf.* la thèse d'Abir Mhenni), l'enrôlement comporte au minimum cinq saisies du mot de passe (le plus souvent, au moins dix), ce qui constitue une

réelle contrainte pour l'utilisateur. L'approche développée par Abir repose sur une seule saisie du mot de passe pour l'enrôlement de l'utilisateur, ce qui présente les avantages suivants :

- une simplicité accrue
- une utilisabilité garantie
- un temps de calcul réduit
- une prise en compte des contraintes des industriels

Cette capture initiale constitue le modèle de référence au début de l'utilisation du système. Elle sera complétée puis remplacée progressivement au cours du processus d'adaptation qui suit.

Bloc Vérification.
La vérification a recours à l'algorithme des K plus proches voisins (KNN, K Nearest Neighbours) pour la vérification de l'identité de l'utilisateur, méthode qui se révèle très efficace dans l'état de l'art. L'algorithme de KNN est testé avec plusieurs choix de distances. Il s'avère que les quatre plus efficaces sont : les distances classiques (euclidienne, de Hamming, de Manhattan) et une distance statistique définie dans l'article [Cam+09]. Pour améliorer les performances considérées séparément, on va combiner les quatre classifieurs résultants au sein d'un vote pondéré, selon l'équation (1.3). Les poids $(\alpha, \beta, \gamma, \delta)$ sont optimisés par algorithme génétique. Le résultat de ce vote pour l'utilisateur j est noté $Score_j$.

$$Score_j = \alpha \times \mathrm{D_{STAT}} + \beta \times \mathrm{D_{HAMM}} + \gamma \times \mathrm{D_{EUCL}} + \delta \times \mathrm{D_{MANH}} \tag{1.3}$$

Ce score personnalisé est alors comparé au seuil de vérification, qui va être adapté pour chaque utilisateur, au fil du temps, comme on va le voir ci-dessous. La fonction de coût de l'algorithme génétique est la minimisation du HTER (Half Total Error Rate, une approximation de l'EER). Les autres paramètres de l'algorithme génétique, ainsi que les autres détails de la procédure sont décrits dans la thèse d'Abir, et dans l'article [Mhe+19b].
Si la requête est acceptée par le système, la phase suivante s'enclenche.

Bloc Adaptation du modèle.
Dans cette phase, les deux seuils pour la vérification et l'adaptation sont mis à jour (*first adapted threshold* et *second adapted threshold*, *cf.* fig. 1.10), pour chaque utilisateur et à chaque session, contrairement à la plupart des approches, qui définissent soit un seuil individualisé, soit un seuil variable. Le critère d'adaptation est un double seuillage. Des seuils individuels et variables au cours du temps sont définis pour chaque individu. Pour l'utilisateur j, le seuil de décision varie selon la formule (1.4), où μ_j et σ_j sont la moyenne et l'écart-type calculés à partir de tous les échantillons de la galerie.

$$T_j^{i+1} = T_j^i - e^{-\frac{\mu_j}{\sigma_j}} \tag{1.4}$$

Le seuil décroît donc progressivement, sa valeur initiale est fixée pour un EER proche de 3%, par conséquent le seuil initial est assez tolérant, tout en garantissant le rejet des imposteurs. Puisque ce seuil est assez haut, les sessions suivantes vont permettre de capturer beaucoup de variabilité, puis de moins en moins, au fil du temps, à mesure que le seuil va diminuer très lentement. L'adaptation de ce seuil de décision suit la progression naturelle du comportement de l'utilisateur, qui gagne en maîtrise dans la saisie de son mot de passe.
Concernant le seuil d'adaptation, il suit aussi la formule (1.4). A chaque session, huit captures sont présentées au système : cinq données de l'utilisateur légitime, et trois données d'imposteurs. Pour la première session, les cinq données authentiques sont présentées, puis les données d'imposteurs. Pour les sessions suivantes, les données d'imposteurs sont présentées dans un ordre aléatoire. La galerie, qui constitue le modèle de référence, est complétée (on rappelle qu'on commence avec un seul échantillon d'enrôlement) jusqu'à comporter dix échantillons. On

parle de processus de *growing window*. Ensuite, lorsqu'un nouvel échantillon est accepté, il vient remplacer le plus ancien échantillon dans la galerie. On parle de processus de *sliding window*. La figure 1.11 illustre le double mécanisme de définition de la galerie.

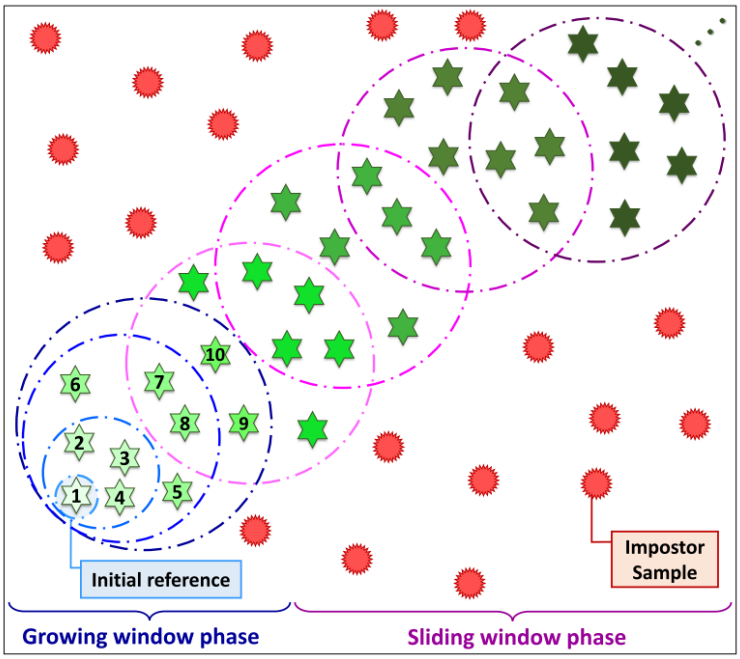


FIGURE 1.11 – Les effets du mécanisme double sur la galerie

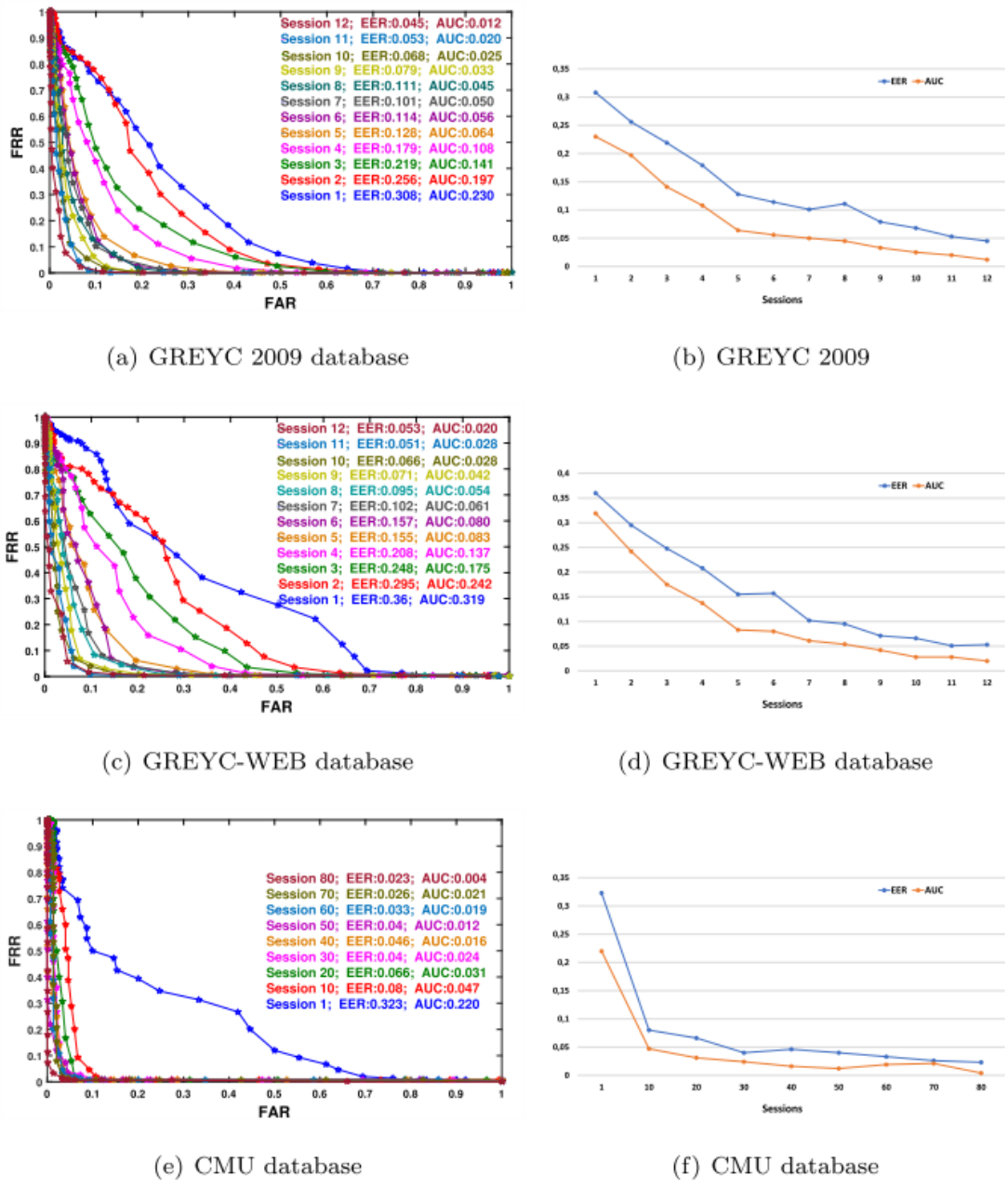
La figure 1.12 présente l'évolution des performances (EER sur la courbe ROC, colonne de gauche et AUC – Area Under Curve –, colonne de droite) de l'approche d'Abir Mhenni – enrôlement unique + mécanisme double d'adaptation du modèle –, performances testées sur trois bases de données classiques de DDF (base CMU [KM10], bases GREYC [GER12]).

Adaptation personnalisée et zoo de Doddington

Une deuxième contribution de la thèse d'Abir Mhenni repose sur l'utilisation de la ménagerie de Doddington, exposée dans l'article de référence [Dod+98]. Il s'agit d'une classification des utilisateurs selon plusieurs profils :

- *les moutons* : les utilisateurs faciles à reconnaître, caractérisés par un FNMR élevé ;
- *les chèvres* : les utilisateurs particulièrement difficiles à reconnaître, qui génèrent une augmentation du FNMR ; cette catégorie peut ne pas être présente dans une base de données ;
- *les agneaux* : les utilisateurs faciles à imiter, qui tendent à augmenter le FMR ;
- *les loups* : les utilisateurs qui peuvent facilement imiter les autres (particulièrement les *agneaux*), qui tendent également à augmenter le FMR.

Doddington *et al.* ont validé le modèle de leur ménagerie sur la reconnaissance de la voix. Ils ont rapporté une corrélation entre la classe des *agneaux* et celle des *loups*, cohérente avec la symétrie présente dans l'algorithme d'appariement. Tester si un utilisateur est un *agneau* nécessite des données d'enrôlement ; tester si un utilisateur est un *loup* nécessite des données de vérification. Lors des tests pour le calcul du FMR, un maximum de données est utilisé, c'est pourquoi les tests croisés tendent à ne considérer qu'une classe pour les agneaux et les loups, même si, intrinsèquement, ce sont deux classes fondamentalement distinctes. Le modèle de la



(a) GREYC 2009 database

(b) GREYC 2009

(c) GREYC-WEB database

(d) GREYC-WEB database

(e) CMU database

(f) CMU database

FIGURE 1.12 – Evolution des performances au fil des sessions (colonne de gauche : courbes ROC, colonne de droite : AUC)

ménagerie a été appliqué à d'autres modalités, à savoir le visage, l'empreinte digitale, l'iris ou encore la signature.

Houmani *et al.* ont proposé dans l'article [HG16] une étude de la ménagerie de Doddington pour la vérification de signature, reposant sur deux mesures de qualité : l'entropie personnelle et l'entropie relative. L'entropie personnelle mesure la complexité et la variabilité de la signature, elle est calculée à partir des données des utilisateurs légitimes uniquement : elle permet de distinguer les *moutons* des *chèvres*. La figure 1.14 illustre l'évolution de l'entropie personnelle de certains utilisateurs présents dans la base WEBGREYC. L'entropie relative est elle aussi une mesure de qualité, calculée à partir de toutes les données (utilisateurs légitimes et imposteurs),

qui indique le niveau de difficulté de l'attaque d'une signature ; elle caractérise la classe des *agneaux*.
La figure 1.13 illustre la répartition des différentes catégories du zoo, en fonction des taux de FMR et FNMR (à gauche, approche initiale de Doddington *et al.* [Dod+98]) ou de l'entropie personnelle et de l'entropie personnelle relative (à droite, selon l'approche développée dans l'article [HG16]).

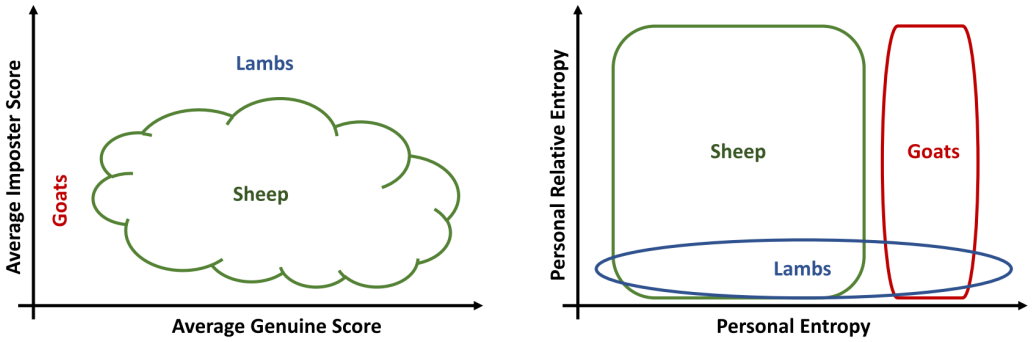


FIGURE 1.13 – Répartition des animaux du zoo de Doddington

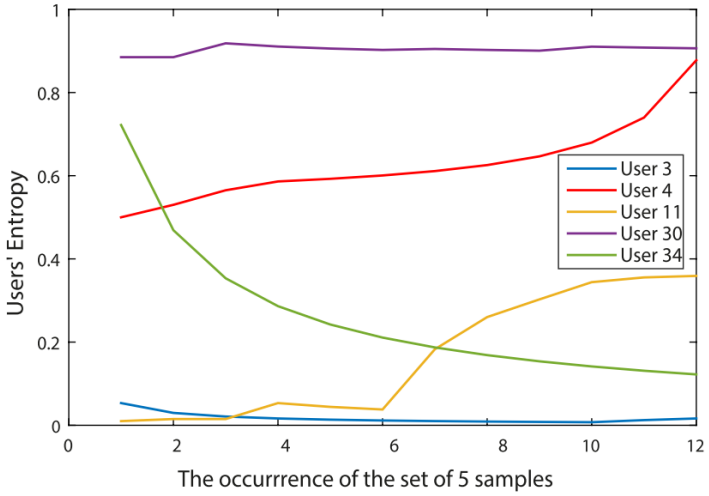


FIGURE 1.14 – Evolution de l'entropie personnelle au fil des sessions

On peut remarquer que la reconnaissance des *moutons* et des *chèvres* ne repose que sur l'étude des scores légitimes (faibles pour les *chèvres*), tandis que la reconnaissance des *agneaux* et des *loups* ne repose que sur l'étude des scores imposteurs (hauts pour les *agneaux* et les *loups*). Dans l'article [YD07], Yager et Dunstone ont complété la ménagerie de Doddington avec quatre nouvelles catégories, grâce à une étude mettant en relation les scores imposteurs et les scores légitimes. Cet article introduit les définitions suivantes. Soit $\mathcal{P}$ l'ensemble des utilisateurs (ou population) et $\mathcal{S}$ l'ensemble des scores de comparaison. Pour chaque couple d'utilisateurs $j, k \in \mathcal{P}$, on note $S(j,k) \subset \mathcal{S}$ l'ensemble contenant les scores de comparaison entre les échantillons de l'utilisateur j avec un modèle de référence de l'utilisateur k. L'ensemble G_k des scores légitimes de l'utilisateur k est défini par $G_k = S(k,k)$ et l'ensemble I_k des scores imposteurs pour le même utilisateur correspond à $I_k = S(j,k) \bigcup S(k,j)$, pour $j \neq k$.
Pour définir les nouvelles catégories d'animaux, Yager *et al.* utilisent les caractérisations sui-

vantes. Soit $\mathcal{G}$ l'ensemble des mesures de performance légitime moyenne (average genuine performance measures) : $\mathcal{G} = \bigcup_{k \in \mathcal{P}} g_k$. On trie les utilisateurs $k \in \mathcal{P}$ dans l'ordre croissant des valeurs statistiques de performance légitime g_k. On note $\mathcal{G}_H$ le sous-ensemble de $\mathcal{P}$ contenant les 25% ayant les scores les plus élevés (High) dans $\mathcal{G}$, et $\mathcal{G}_L$ celui contenant les 25% les plus bas (Low). De la même façon, on définit $\mathcal{I}_H$ et $\mathcal{I}_L$ pour les scores imposteurs.
A partir de ces différentes notations, Yager et Dunstone proposent les nouveaux animaux suivants [YD07] :

- *les caméléons* = les utilisateurs assez similaires aux autres, obtenant par conséquent des scores élevés pour les comparaisons légitimes ou imposteurs ; ils sont dans l'ensemble $\mathcal{G}_H \cap \mathcal{I}_H$;
- *les fantômes* = les utilisateurs souvent rejetés par le système, présentant des scores (légitimes ou imposteurs) assez bas, dans l'ensemble $\mathcal{G}_L \cap \mathcal{I}_L$;
- *les colombes* les « meilleurs » utilisateurs, faciles à reconnaître et difficiles à attaquer ; ils sont dans l'ensemble $\mathcal{G}_H \cap \mathcal{I}_L$;
- *les vers* : les « pires » utilisateurs, difficiles à reconnaître et faciles à attaquer ; ils sont dans l'ensemble $\mathcal{G}_L \cap \mathcal{I}_H$.

La figure 1.15 illustre la répartition de la ménagerie complète, en termes de scores de comparaison et en termes d'entropie personnelle/personnelle relative.

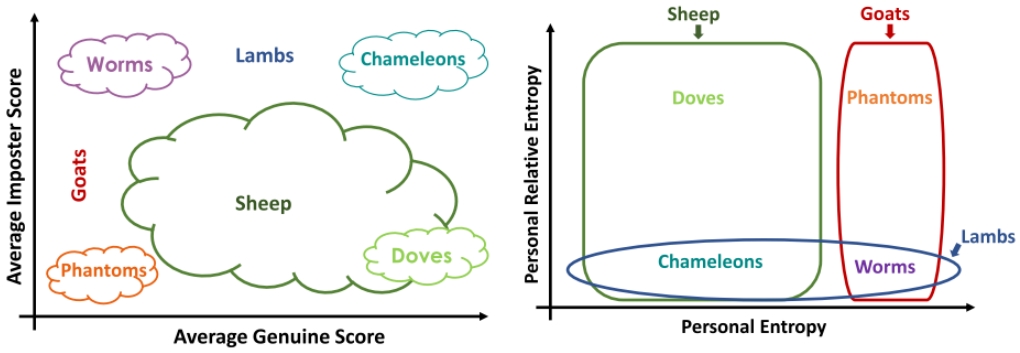


FIGURE 1.15 – Répartition complète des animaux du zoo de Doddington [YD07]

Dans sa thèse, Abir Mhenni a considéré l'approche de Houmani *et al.* [HG16] pour personnaliser l'adaptation du modèle en fonction du profil de l'utilisateur, le profil correspondant à une espèce de la ménagerie. Si on se réfère à la figure 1.14, on constate que l'entropie personnelle de l'utilisateur 34 décroît au fil du temps, tandis que celle des utilisateurs 4 et 11 augmente, et celle des utilisateurs 3 et 30 reste stable. Cette mesure de la difficulté à attaquer les données d'un utilisateur sera exploitée pour assigner une catégorie d'animaux à chaque utilisateur présent dans la base de données. Deux contributions successives et complémentaires reposent sur la ménagerie de Doddington pour la mise à jour du modèle biométrique : la première considère trois classes (*mouton*, *chèvre*, *agneau*), la seconde, la ménagerie complète (soit sept catégories d'animaux). La catégorie *loup* n'est jamais considérée, car on ne souhaite pas se placer du point de vue des imposteurs.

• *Première approche : adaptation avec trois catégories d'animaux*

Le système d'authentification adaptatif reprend le principe de la contribution précédente, avec

un enrôlement simple, comportant une capture initiale unique, dans un souci d'utilisabilité (*cf.* le chapitre suivant). Tout au long de l'utilisation du système, le modèle de référence est enrichi par un processus additif (ou *growing window*), puis mis à jour par un processus de remplacement (ou *sliding window*). Durant la première phase, les nouveaux échantillons sont ajoutés à la référence jusqu'à atteindre la taille maximale, qui dépendra de la classe d'animaux. Les utilisateurs sont alors caractérisés par l'évolution de la taille de leur galerie au fil des sessions. Si cette taille augmente lentement, il s'agit d'utilisateurs difficiles à reconnaître, donc appartenant à la classe *chèvres*. Si au contraire, la galerie se remplit rapidement, il s'agit d'utilisateurs faciles à reconnaître, donc appartenant à la classe *moutons*.

Comme dans la contribution précédente, quatre distances sont calculées entre les échantillons de la galerie et la nouvelle capture. Un score est calculé pour chaque distance sélectionnée (Manhattan, Hamming, Euclidienne, statistique) grâce à l'algorithme des K plus proches voisins. Ces scores sont ensuite combinés dans un vote pondéré, dont les poids sont optimisés par algorithme génétique à la fin de chaque session.
Une session correspond à un flux de données réparties comme suit : pour chaque utilisateur, on présente huit captures au système, réparties en cinq données légitimes, et trois données d'imposteurs. Plusieurs scénarios sont considérés dans la thèse d'Abir concernant l'ordre de présentation.
Remarque : comme chaque session comporte la présentation de huit captures, il peut arriver que la taille maximale pour la galerie soit dépassée. Cela ne peut concerner que les cas où un utilisateur de type *chèvre* change de catégorie pour devenir *mouton*. Dans ce cas, les échantillons les moins utiles de la galerie sont retirés (processus *least frequently used*).
Ensuite, la mise à jour du modèle va être spécifique pour chaque catégorie d'animaux. On considère ici la situation où la galerie est complète : chaque utilisateur a été classifié en *mouton* ou *chèvre*. A ce stade, c'est l'entropie (personnelle et relative) qui permet de poursuivre la classification. Les formules sont détaillées dans l'article [HG16]. Le calcul de l'entropie personnelle, qui mesure la variabilité *intra*, assigne à l'utilisateur soit la catégorie *mouton* (entropie faible), soit la catégorie *chèvre* (entropie élevée). Ensuite, la vulnérabilité aux attaques, via le calcul de l'entropie relative, permet de reconnaître les *agneaux* (entropie relative faible).
A la fin de chaque session, lorsque la répartition des utilisateurs dans les différentes catégories est terminée, les poids impliqués dans le vote des quatre classifieurs basés sur les K plus proches voisins sont optimisés par algorithme génétique pour suivre les variations de chaque catégorie (la fonction de coût est la minimisation du FMR et du FNMR). C'est à ce moment également que les différents seuils sont mis à jour, avec une stratégie différente selon la catégorie :

- catégorie *moutons* : la taille de la galerie est dix échantillons, et l'évolution des seuils de vérification et d'adaptation est donnée par la formule (1.4) ;
- catégorie *chèvres* : la taille de la galerie est quinze échantillons (pour conserver plus de variabilité *intra*), et l'évolution des seuils de vérification et d'adaptation est donnée par la formule (1.4) ;
- catégorie *agneaux* : la taille de la galerie est dix échantillons, et les seuils de vérification et d'adaptation sont plus stricts (car les agneaux sont plus sensibles aux attaques par imitation), selon la formule (1.5) :

$$T_j^{i+1} = T_j^i - e^{-\frac{\mu_j}{2\sigma_j}} \tag{1.5}$$

Les résultats obtenus seront présentés plus loin dans le tableau 1.5 avec ceux de l'approche plus complète, qui comporte sept catégories d'animaux.

• *Seconde approche : adaptation avec sept catégories d'animaux*

Le système global est représenté à la figure 1.16.

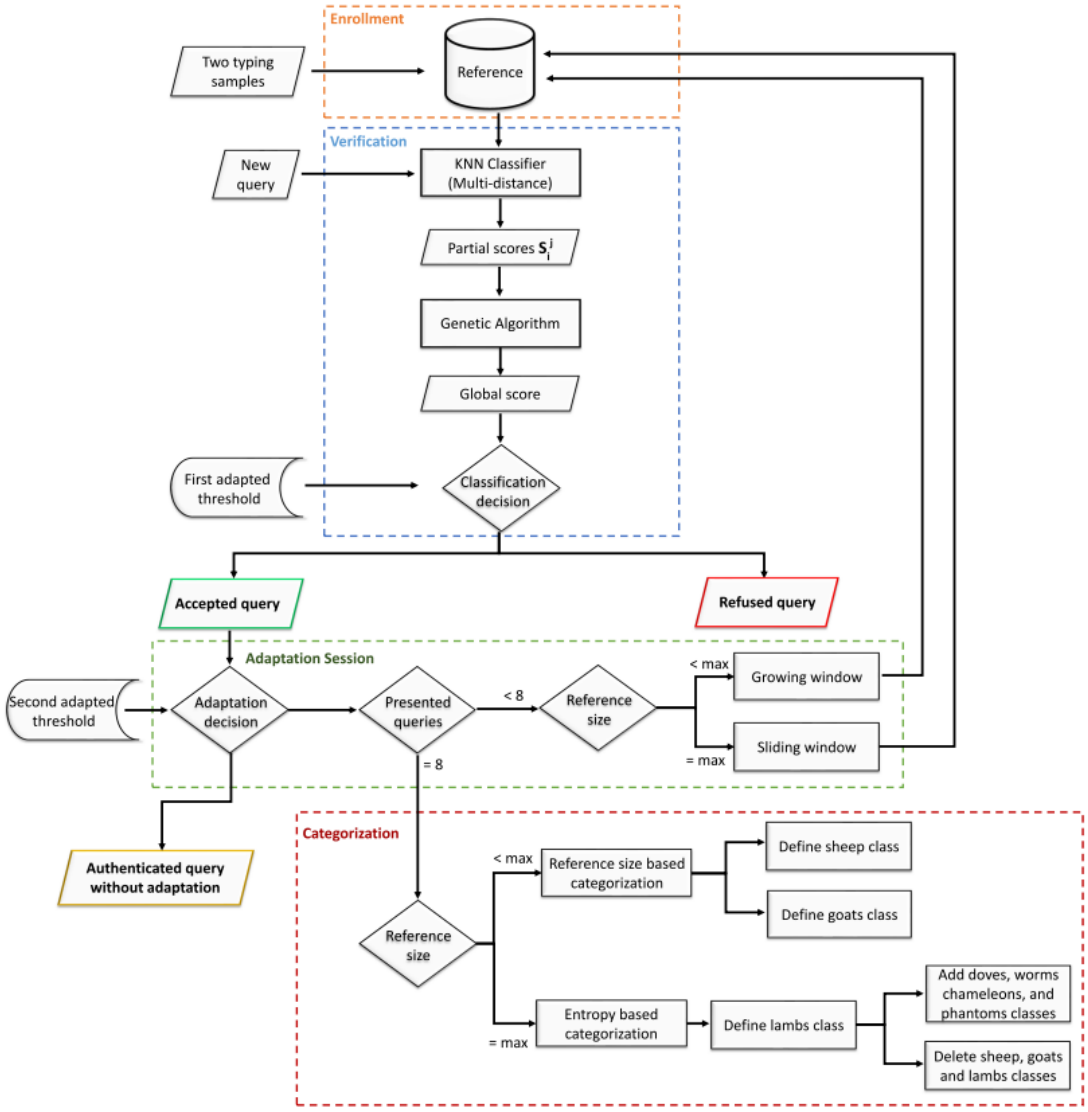


FIGURE 1.16 – Mise à jour du modèle et zoo de Doddington

La première évolution concerne l'enrôlement : deux échantillons sont requis (au lieu d'un seul), en accord avec les stratégies actuelles de définition de mots de passe, qui imposent de saisir un nouveau mot de passe deux fois de suite. Comme dans l'approche précédente, lors de la phase de complétion de la galerie, les utilisateurs sont placés, soit dans la catégorie *moutons*, soit dans la catégorie *chèvres* (voir les détails dans la thèse d'Abir Mhenni [Mhe19], page 134). Lors de la phase suivante, *i.e.* le remplacement des échantillons par fenêtre glissante, les *agneaux* sont repérés comme précédemment. Ensuite, ces trois catégories sont remplacées par les catégories des *vers*, des *colombes*, des *caméléons* et des *fantômes*. La stratégie d'adaptation pour chaque catégorie diffère des autres selon deux critères, elle est résumée dans le tableau 1.4 :

- la taille de la galerie : dix par défaut, quinze pour la catégorie des *vers* qui, comme les

chèvres, présentent plus de variabilité intra, et vingt pour les fantômes, qui sont particulièrement difficiles à décrire ;
- la formule de mise à jour des seuils de vérification et d'adaptation : soit la formule (1.4) (Adapted threshold), soit la formule (1.5) (Stricter threshold), pour éliminer les fausses acceptions pour les catégories sujettes aux attaques (vers, caméléons et fantômes).

User category	Reference size	Thresholds
Sheep	10	Adapted thresholds
Goats	15	Adapted thresholds
Lambs	10	Stricter thresholds
Worms	15	Stricter thresholds
Chameleons	10	Stricter thresholds
Doves	10	Adapted thresholds
Phantoms	20	Stricter thresholds

Tableau 1.4 – Mise à jour personnalisée pour chaque catégorie d'animaux

• *Résultats obtenus*

On rappelle que chaque session repose sur la présentation de huit échantillons pour chaque utilisateur : cinq échantillons légitimes et trois imposteurs. Le tableau 1.5 présente les résultats obtenus sur la base de données WEBGREYC [GER12], pour huit sessions d'adaptation, pour trois méthodes : l'adaptation sans l'approche du zoo de Doddington, avec l'approche comportant trois catégories d'animaux puis sept catégories. Les deux critères d'évaluation sont l'EER (Equal Error Rate), et l'AUC (Area Under Curve). On constate l'efficacité des approches développées par Abir Mhenni dans sa thèse, qui lui ont permis d'obtenir la récompense « Best Full Paper Award » à la conférence Cyberworlds 2018 [Mhe+18].

Adaptation strategy	EER	AUC
Without Doddington menagerie	5.3%	0.02
Biometric menagerie based on 3 classes	0.8 %	0.003
Biometric menagerie based on 7 classes	0.2%	0.002

Tableau 1.5 – Comparaison des résultats de la mise à jour de modèle avec le zoo de Doddington [Mhe19]

D'autres résultats ont été obtenus, que je ne détaillerai pas dans ce manuscrit :
- une étude a été menée sur l'impact des tailles maximales des galeries – tailles différentes en fonction de la catégorie d'animaux –, qui a montré que les tailles retenues au départ sont optimales ;
- différents scénarios ont été testés pour la présentation des données (légitimes ou imposteurs) : les scénarios plus réalistes, où les données d'imposteurs ne sont pas présentes au début fournissent naturellement les meilleurs résultats ;
- une étude a été présentée dans l'article [Mhe+19a], sur les vulnérabilités des systèmes de mise à jour du modèle biométrique vis-à-vis de différentes attaques.

Perspectives

La personnalisation de la stratégie de mise à jour du modèle biométrique qui s'appuie sur la ménagerie de Doddington et ses extensions est particulièrement efficace pour améliorer les performances de la reconnaissance par dynamique de frappe au clavier. Il serait intéressant d'étudier si ce gain en performance est aussi significatif pour d'autres modalités comportementales. Abir Mhenni a également proposé une étude de sécurité des systèmes de biométrie adaptative, en explorant certaines vulnérabilités face à des attaques bien choisies dans l'article [Mhe+19a].

4 Conclusion

Ce premier chapitre a abordé les aspects de sécurité liés aux systèmes de reconnaissance biométrique. Après quelques rappels historiques, les contraintes réglementaires puis les analyses de risques ont été abordées, en lien notamment avec les besoins de sécurité des données biométriques, qui constituent une classe de données personnelles particulièrement sensibles. Ce point sera approfondi dans le chapitre 3. La seconde partie de ce chapitre est consacrée aux contributions extraites des thèses que j'ai encadrées : celle de Syed Idrus Syed Zulkarnain, sur la biométrie douce ; puis celle d'Abir Mhenni, sur la mise à jour de modèle. Ces deux thématiques – biométrie douce et mise à jour de modèle –, offrent des possibilité de renforcement de la sécurité des systèmes d'authentification biométrique par reconnaissance de la dynamique de frappe au clavier par une amélioration de leurs performances.
De récents articles de Lovisotto *et al.* [LEM20] et Orrù *et al.* [Orr+20], proposent des visions radicalement différentes sur la mise à jour de modèle appliquée à la reconnaissance faciale. Dans l'article [LEM20], les auteurs démontrent qu'il est possible d'exploiter le processus d'adaptation pour usurper l'identité d'utilisateurs légitimes, imperceptiblement et sans effort. Cette *attaque par empoisonnement* repose sur la présentation astucieuse d'une suite d'échantillons imposteurs qui vont combler petit à petit la distance avec les échantillons authentiques. En outre, aucun accès au capteur n'est requis, ni aucune connaissance sur les paramètres du système (seuil de décision ou données authentiques). D'un autre côté, les auteurs de l'article [Orr+20] démontrent l'intérêt des techniques de mise à jour du modèle, dans un contexte d'apprentissage profond, pour compenser la variabilité qui affecte la reconnaissance faciale, lorsque plusieurs mois voire plusieurs années séparent les données d'enrôlement et les données présentées au système.
Les mécanismes qui permettent de renforcer la sécurité d'un système d'authentification biométrique, que ce soit la biométrie douce, la mise à jour du modèle, ou autres, sont aujourd'hui intégrés dans l'authentication basée sur les risques (*risk-based authentication*). Celle-ci, contrairement à une authentification statique, à un facteur, repose sur plusieurs éléments, parmi lesquels : une analyse des risques, une authentification multi-facteurs, continue et des informations collectées sur le smartphone ou un objet connecté destiné à réaliser des transactions électroniques. Cette tendance à la sécurité à tout prix pose cependant de nombreuses questions éthiques, quant à la collecte massive de données personnelles, collecte qui se révèle loin d'être transparente pour l'utilisateur. Dans l'article [WID19], Wiefling *et al.* soulèvent quelques limitations et questions éthiques en conclusion, même si le domaine financier notamment pousse de telles tendances à se développer.
C'est peut-être en réaction à un tel mépris pour ce qui constitue l'essence même de notre identité numérique que certaines recherches proposent d'autres voies. Ainsi Mirjalili *et al.* présentent un mécanisme additionnel pour la reconnaissance faciale [MRR18] : le but de ce mécanisme, qui repose sur des techniques d'augmentation des données (*data augmentation*), est de per-

turber les images représentant les visages, suffisamment pour contrecarrer toute tentative de reconnaissance du sexe de l'utilisateur. La problématique posée par ces travaux rejoint les préoccupations européennes exprimées dans le Règlement Général sur la Protection des Données : quid de la finalité, la proportionnalité, la minimisation des données nécessaires à un traitement précis ? Ces questionnements se retrouvent au cœur de mon projet de recherche, exposé dans le chapitre 4.
Dans le chapitre suivant, en lien avec les besoins de sécurité accrus induits par les nouveaux usages et services numériques, seront abordés les aspects d'utilisabilité des systèmes de reconnaissance biométrique.

5 Références du chapitre 1

Bibliographie

[El-+12] M. EL-ABED, R. GIOT, B. HEMERY, J.-J. SCHWARTZMANN et C. ROSENBERGER. « Towards the Security Evaluation of Biometric Authentication Systems ». In : *IACSIT International Journal of Engineering and Technology* (2012).

[Abo11] I. ABOUT. « L'identification biométrique. Champs, acteurs, enjeux et controverses ». In : sous la dir. d'A. CEYHAN et P. PIAZZA. Maison des sciences de l'homme, 2011. Chap. Classer le corps : l'anthropométrie judiciaire et ses alternatives, 1880-1930, p. 39-62.

[BC05] N. BARTLOW et B. CUKIC. *The vulnerabilities of biometric systems - an integrated look and old and new ideas*. Rapp. tech. West Virginia University, 2005.

[Brö06] A. BRÖMME. « A Risk Analysis Approach for Biometric Authentication Technology ». In : *International Journal of Network Security* (2006).

[Cam+09] P. CAMPISI, E. MAIORANA, M. BOSCO et A. NERI. « User authentication using keystroke dynamics for cellular phones ». In : *IET Signal Processing* (2009).

[Cam13] CAMPISI, P. (EDITOR). *Security and Privacy in Biometrics*. Springer-Verlag, 2013.

[Den11] V. DENIS. « L'identification biométrique. Champs, acteurs, enjeux et controverses ». In : sous la dir. d'A. CEYHAN et P. PIAZZA. Maison des sciences de l'homme, 2011. Chap. Identifier le corps avant la biométrie aux XIV$^{\text{ème}}$- XIX$^{\text{ème}}$siècles, p. 25-38.

[DH17] L. M. DINCA et G. P. HANCKE. « The Fall of One, the Rise of Many: A Survey on Multi-Biometric Fusion Methods ». In : *IEEE Access* (2017).

[Dod+98] G. DODDINGTON, W. LIGGETT, A. MARTIN, M. PRZYBOCKI et D. REYNOLDS. *Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation*. Rapp. tech. NATIONAL INST OF STANDARDS et TECHNOLOGY GAITHERSBURG MD, 1998.

[ENI19] ENISA. *Standardisation in support of the cybersecurity certification. Recommendations for European standardisation in relation to the Cybersecurity Act.* 2019.

[ELM11] C. EPP, M. LIPPOLD et R. MANDRYK. « Identifying emotional states using keystroke dynamics ». In : *Conference on Human Factors in Computing Systems*. 2011.

[Fer13] M. B. FERNÁNDEZ SAAVEDRA. « Evaluation methodologies for security testing biometric systems beyond technological evaluation ». Thèse de doct. Universidad Carlos III de Madrid, 2013.

[Gio12] R. GIOT. « Contributions à la dynamique de frappe au clavier : multibiométrie, biométrie douce et mise à jour de la référence ». Thèse de doct. Université de Caen Normandie, 2012.

[GER12] R. GIOT, M. EL-ABED et C. ROSENBERGER. « Web-Based Benchmark for Keystroke Dynamics Biometric Systems: A Statistical Analysis ». In : *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. 2012.

[Gio+12] R. GIOT, B. HEMERY, E. CHERRIER et C. ROSENBERGER. « La multibiométrie ». In : *Traitement du signal et de l'image pour la biométrie*. Hermès, 2012, Chapitre 9.

[HG16] N. HOUMANI et S. GARCIA-SALICETTI. « On Hunting Animals of the Biometric Menagerie for Online Signature ». In : *PLOS ONE* (2016).

[Ica09] S. ICARD. *La fiche-numéro et le registre digital*. Archive d'anthropologie criminelle (tome 24). 1909.

[JRN11] A. K. JAIN, A. A. ROSS et NANDAKUMAR. *Introduction to biometrics*. Springer Publishing Company, Incorporated, 2011.

[JNN08] A. K. JAIN, K. NANDAKUMAR et A. NAGAR. « Biometric Template Security ». In : *EURASIP Journal on Advances in Signal Processing* (2008).

[JMD20] M. JOSHI, B. MAZUMDAR et S. DEY. « A comprehensive security analysis of match-in-database fingerprint biometric system ». In : *Pattern Recognition Letters* (2020).

[KM10] K. KILLOURHY et R. MAXION. « Why Did My Detector Do That?! Predicting Keystroke-Dynamics Error Rates ». In : *Recent Advances in Intrusion Detection*. 2010.

[Kin12] E. KINDT. « The Processing of Biometric Data - A Comparative Legal Analysis with a focus on the Proportionality Principle and Recommendations for a Legal Framework ». Thèse de doct. Belgique : Université Catholique de Louvain, 2012.

[LEM20] G. LOVISOTTO, S. EBERZ et I. MARTINOVIC. « Biometric Backdoors: A Poisoning Attack Against Unsupervised Template Updating ». In : *IEEE European Symposium on Security and Privacy*. 2020.

[Mhe+19a] A. MHENNI, D. MIGDAL, E. CHERRIER, C. ROSENBERGER et N. ESSOUKRI BEN AMARA. « Vulnerability of Adaptive Strategies of Keystroke Dynamics Based Authentication Against Different Attack Types ». In : *2019 International Conference on Cyberworlds (CW)*. 2019.

[Mhe19] A. MHENNI. « Contribution to the biometric template update: Application to keystroke dynamics modality ». Thèse de doct. National Engineering School of Tunis, 2019.

[Mhe+19b] A. MHENNI, E. CHERRIER, C. ROSENBERGER et N. E. B. AMARA. « Single Enrollment for Keystroke Dynamics with Adaptive Template Update ». In : *Computers & Security* (2019).

[Mhe+18] A. MHENNI, E. CHERRIER, C. ROSENBERGER et N. ESSOUKRI BEN AMARA. « User Dependent Template Update for Keystroke Dynamics Recognition ». In : *Cyberworlds*. Singapore, 2018.

[MRR18] V. MIRJALILI, S. RASCHKA et A. ROSS. « Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers ». In : *IEEE 9th International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. 2018.

[Orr+20] G. ORRÙ, M. MICHELETTO, J. FIÉRREZ et G. L. MARCIALIS. « Are Adaptive Face Recognition Systems still Necessary? Experiments on the APE Dataset ». In : *4th IEEE International Conference on Image Processing, Applications and Systems, IPAS 2020, Virtual Event, Italy, December 9-11, 2020*. 2020.

[Pis+19] P. H. PISANI, A. MHENNI, R. GIOT, E. CHERRIER, A. C. P. L. F. de CARVALHO, N. E. B. AMARA et C. ROSENBERGER. « Adaptive Biometric Systems: Review and Perspectives ». In : *ACM Computing Surveys* (2019).

[RCB01] N. K. RATHA, J. H. CONNELL et R. M. BOLLE. « Enhancing security and privacy in biometrics-based authentication systems ». In : *IBM Systems Journal* (2001).

[Rob07] C. ROBERTS. « Biometric attack vectors and defences ». In : *Computers & Security* (2007).

[San+17] B. SANTOSO, H. WIJAYANTO, K. A. NOTODIPUTRO et B. SARTONO. « Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review ». In : *IOP Conference Series: Earth and Environmental Science* (2017).

[SSR19] M. SINGH, R. SINGH et A. ROSS. « A comprehensive overview of biometric fusion ». In : *Information Fusion* (2019).

[Sye14] S. Z. SYED IDRUS. « Soft biometrics for keystroke dynamics ». Thèse de doct. Université de Caen Normandie, 2014.

[Sye+14] S. Z. SYED IDRUS, E. CHERRIER, C. ROSENBERGER et P. BOURS. « Soft Biometrics for Keystroke Dynamics: Profiling Individuals While Typing Passwords ». In : *Computers & Security* 45 (2014), p. 147-155.

[Sye+13] S. Z. SYED IDRUS, E. CHERRIER, C. ROSENBERGER et J.-J. SCHWARTZMANN. « A Review on Authentication Methods ». In : *Australian Journal of Basic and Applied Sciences* 7.5 (mars 2013), p. 95-107.

[Tin02] K. M. TING. « An instance-weighting method to induce cost-sensitive trees ». In : *IEEE Transactions on Knowledge and Data Engineering* (2002).

[VCC99] K. VEROPOULOS, C. CAMPBELL et N. CRISTIANINI. « Controlling the Sensitivity of Support Vector Machines ». In : *Proceedings of International Joint Conference Artificial Intelligence* (1999).

[Way96] J. L. WAYMAN. « Technical Testing and Evaluation of Biometric Identification Devices ». In : *Biometrics: Personal Identification in Networked Society*. Sous la dir. d'A. K. JAIN, R. BOLLE et S. PANKANTI. Springer US, 1996.

[WID19] S. WIEFLING, L. L. IACONO et M. DÜRMUTH. « Is this really you? An empirical study on risk-based authentication applied in the wild ». In : *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer. 2019.

[YD07] N. YAGER et T. DUNSTONE. « Worms, Chameleons, Phantoms and Doves: New Additions to the Biometric Menagerie ». In : *IEEE Workshop on Automatic Identification Advanced Technologies*. 2007.

CHAPITRE

2

Biométrie et utilisabilité

Le simple est toujours faux. Ce qui ne l'est pas est inutilisable.

Mauvaises pensées et autres, Paul Valéry, 1941

Selon le standard ISO/IEC 9241-11:2018 *Ergonomie de l'interaction homme-système*, l'utilisabilité est un concept général souvent employé dans deux acceptions : la *facilité d'utilisation* ou la *convivialité*.

L'authentification biométrique possède la capacité immédiate d'améliorer l'utilisabilité des smartphones et autres objets connectés. En effet, la possibilité d'accéder à une application, un service, de payer, etc., juste par une pression du doigt ou un simple regard en direction d'un capteur représente une simplification évidente par rapport à la saisie d'un mot de passe complexe, d'un code PIN ou la présentation d'un badge. Cependant, la règle générale veut que, en dehors des performances, point d'intérêt : par conséquent, le point de vue de l'utilisateur est rarement pris en compte dans la conception d'un système d'authentification biométrique.

Cette citation du NIST en est l'illustration :
« *In order to improve the usability of biometric systems, it is critical to take a holistic approach that considers the needs of users as well as the entire experience users will have with a system, including the hardware, software and instructional design of a system. Adopting a user-centric view of the biometric process is not only beneficial to the end users, but a user-centric view can also help to improve the performance and effectiveness of a system.* [1] »

Au-delà de la facilité d'usage, l'étude de l'utilisabilité implique de changer de point de vue pour remettre l'individu au centre du processus, et de s'appuyer sur des domaines complémentaires, tels que la psychologie, le design, l'ergonomie, etc. C'est tout l'objet de ce chapitre.

1. https://www.nist.gov/programs-projects/usability-biometric-systems

1 Le besoin d'utilisabilité dans les solutions d'authentification biométrique

Dans l'article [KSE10], Kulula *et al.* expliquent que, pour que le déploiement d'un système biométrique soit un succès, il faut prendre en considération la façon dont les utilisateurs interagissent avec ce système. Tout manquement à cette recommandation mène inévitablement à une dégradation de l'utilisation optimale du capteur biométrique, et par conséquent à une dégradation des performances, sous forme d'échec lors de l'acquisition, lors de l'enrôlement, ou sous forme de faux rejets supplémentaires. L'article de Blanco-Gonzalo *et al.* [Bla+13] pousse le raisonnement plus loin : pour tendre vers un système biométrique universel, il faut améliorer la compréhension de l'utilisateur. Cet effort inclut l'étude des interactions avec le capteur et le système tout entier, tant au niveau physique que cognitif : le but est de favoriser un véritable apprentissage des techniques d'authentification biométrique en termes d'utilisation, voire un transfert de connaissance d'une technologie à l'autre par l'utilisateur lui-même. On constate donc que l'utilisabilité est un des facteurs qui influence le plus les performances finales. Elle constitue par conséquent un des aspects incontournables de la conception d'un système biométrique, au même titre que la sécurité et le respect de la vie privée (voir les chapitres 1 et 3).

1.1 Une relation ambiguë entre les systèmes biométriques et les utilisateurs

L'utilisabilité d'une nouvelle technologie a donc un impact direct sur son adoption par les utilisateurs. Cette remarque est d'autant plus vraie pour la biométrie, qui souffre d'un passif lié à des dérives sécuritaires relatées dans la plupart des dystopies. Par exemple dans *Minority Report* [2], où le héros doit changer ses yeux – au sens propre – pour échapper au système de reconnaissance déployé partout dans le but d'arrêter les criminels potentiels avant qu'ils ne commettent leur crime. Ou encore dans le roman *Les Furtifs* [3], où les personnes qui ne veulent pas être tracées (par l'intermédiaire d'une bague) sont marginalisées : « *La bague au doigt, vous êtes tout à fait libres et parfaitement tracés, soumis au régime d'auto-aliénation consentant propre au raffinement du capitalisme cognitif* [4] ». Cependant, depuis quelques années, avec l'essor des smartphones, tablettes et autres objets connectés, la biométrie apparaît comme une technologie simplificatrice, que les utilisateurs utilisent volontiers sous la forme de reconnaissance d'empreinte digitale, de reconnaissance faciale ou vocale, de façon transparente, sans se soucier du caractère sensible des données biométriques stockées et analysées dans ces objets du quotidien. L'ergonomie, la simplicité, la facilité d'utilisation l'emportent indéniablement sur les inquiétudes potentielles quant au respect de la vie privée. Les technologies actuelles, embarquées dans les smartphones et les nouveaux objets connectés, ont rendu caduc l'éternel compromis entre la sécurité – y compris la sécurité des données, donc leur protection et le respect de la vie privée – et l'utilisabilité. Jusqu'à récemment, plus de sécurité signifiait invariablement moins de convivialité : les mots de passe de plus en plus complexes (à changer très régulièrement), les codes PIN et autres informations à mémoriser, les CAPTCHA, etc., représentent un réel fardeau pour notre cerveau. L'authentification biométrique a permis d'alléger ce fardeau, tout en garantissant un niveau de sécurité élevé. Cependant, et on le verra plus loin au paragraphe 2.4 avec le modèle BioTAM, l'enjeu majeur pour les concepteurs de systèmes biométriques est d'instaurer la confiance des utilisateurs qui n'est pas acquise *a priori*.

2. *Minority Report* de Philip K. Dick, publiée en janvier 1956 dans la revue *Fantastic Universe*
3. Les Furtifs d'Alain Damasio, publié en 2019 aux Éditions La Volte
4. https://lavolte.net/livres/les-furtifs-alain-damasio/

1.2 Les nouveaux usages et les nouvelles réglementations

La biométrie offre des solutions d'authentification à la fois sécurisées (voir le chapitre 1) et respectueuses de la vie privée (voir le chapitre 3). A côté de ces propriétés incontournables, leur utilisabilité est mise en avant à la fois dans la littérature et par les industriels, depuis l'intégration de capteurs biométriques dans les smartphones notamment. Ces trois critères {sécurité + utilisabilité + respect de la vie privée} incitent les acteurs publics et privés à proposer des solutions d'authentification forte, incluant la biométrie, en réponse à de nouvelles réglementations européennes (eIDAS, DSP2, 3D secure2 pour en citer quelques-unes). De nouveaux usages apparaissent donc, qui nécessitent le développement de systèmes biométriques performants.

1.2.1 Identité numérique régalienne française et eIDAS : *Electronic IDentification Authentication and trust Services*

Depuis 2014, le Règlement eIDAS n° 910/2014 définit un socle commun pour tous les échanges électroniques sécurisés entre les citoyens, les entreprises et les autorités publiques au sein de l'Union Européenne. L'Europe met en avant trois niveaux de sécurité – *faible*, *substantiel* et *élevé* – correspondant à des besoins distincts. Or, aujourd'hui la France va seulement disposer d'une solution de niveau *substantiel* ou *élevé*, sous la forme d'une carte d'identité électronique, déployée à partir du 16 mars 2021 [5]. Malgré le système de fédération d'identité *France connect*, qui simplifie l'accès à certains services administratifs en ligne, le niveau de sécurité est encore globalement insuffisant en France pour couvrir l'ensemble des besoins. Pour être en conformité avec le droit européen en 2021, la France a dû se doter d'une identité numérique régalienne répondant aux standards européens de sécurité et interopérable avec celle des autres pays membres. Dans ce but, deux missions avaient été lancées, l'une par le Gouvernement, l'autre par l'Assemblée Nationale. La première était une mission interministérielle de 2018, consacrée au déploiement d'un parcours d'identification numérique sécurisé. La mission d'information parlementaire sur l'identité numérique, quant à elle, s'est intéressée *aux enjeux « d'éthique, de confiance, de sécurité, d'inclusion des citoyens et de protection de leurs droits » soulevés par la mise en œuvre d'une identité numérique régalienne*. [6] Cette mission a fait valoir que « le succès de la diffusion d'une solution d'identité numérique dépend fortement de la capacité des citoyens à s'en saisir pour leurs usages de la vie quotidienne et du secteur privé à privilégier cette solution par rapport aux autres offres proposées par certains acteurs économiques ». Outre les enjeux de protection des données privées et l'accessibilité de la solution d'identité numérique retenue, l'usage de la biométrie est questionné par les travaux de cette mission, à travers une consultation citoyenne qui a été lancée en mars 2020. La solution retenue, à savoir la carte d'identité électronique, va permettre, dans le cadre du Règlement eIDAS, de porter une identité numérique régalienne et de mettre en œuvre une fonction de signature électronique grâce à la puce intégrée dans la carte. L'interopérabilité des enjeux de sécurité informatique est par conséquent renforcée entre les pays européens. Pour simplifier les démarches administratives des citoyens français, cette carte d'identité électronique constitue un moyen pour d'autres acteurs de dériver des identités numériques fédérées dans FranceConnect.
Le chapitre [Cap20] va plus loin et propose une analyse juridique critique et réaliste du règlement eIDAS :
« *Le système d'identification électronique de l'Union européenne se construit peu à peu et la confrontation à la pratique permet de gommer progressivement les aspérités ou les points d'ombre qui*

5. https://www.gouvernement.fr/tout-ce-qu-il-faut-savoir-sur-la-nouvelle-carte-nationale-d-identite

6. http://www2.assemblee-nationale.fr/15/missions-d-information/missions-d-information-communes/identite-numerique/(block)/67684

persistent dans un domaine particulièrement complexe alliant droit et sécurité technique, mais son utilité pratique est incontestable et la confiance du marché en dépend. »

On constate ainsi que la réglementation européenne influence le recours à l'authentification biométrique pour développer, au niveau de chaque état membre, une solution d'identité numérique régalienne. L'accès à des services administratifs dont le niveau de sécurité est estimé « substantiel » ou « élevé » passera donc bientôt par une authentification biométrique en France. C'est le cas également pour les services de paiement, qui sont en constante évolution.

1.2.2 DSP2 : Directive européenne sur les Services de Paiement, version 2

La Directive européenne 2015/2366/UE sur les Services de Paiement (DSP 2) adoptée le 25 novembre 2015 remplace la première Directive 2007/64/CE, dite DSP 1, mise en œuvre par tous les États membres depuis le 1er novembre 2009. Des normes techniques de réglementation relatives à l'authentification forte du client et à des normes ouvertes communes et sécurisées de communication ont été élaborées par l'Autorité bancaire européenne pour compléter la directive, sous la forme du règlement délégué (UE) 2018/389, adopté le 27 novembre 2017. Cette directive a été transposée dans la loi française (loi n° 2018-700 ratifiant l'ordonnance n° 2017-1252 du 9 août 2017). Elle est parue au Journal officiel n° 179 du 5 août 2018 et vise à [7] :

- créer deux nouveaux services de paiement (les services d'initiation de paiement et les services d'information sur les comptes),
- adapter le régime applicable aux établissements de paiement,
- renforcer des prérogatives de l'État membre d'accueil,
- assurer la sécurité et la protection des consommateurs.

L'article 97, paragraphe 1 de cette directive stipule que l'authentification de l'utilisateur doit être *forte*, c'est-à-dire fondée sur deux ou plusieurs facteurs d'authentification, dont voici un rappel :

- *facteur de connaissance* : désigne quelque chose que seul l'utilisateur connaît
- *facteur de possession* : désigne quelque chose que seul l'utilisateur possède
- *facteur d'inhérence* : désigne quelque chose que l'utilisateur est

La DSP2 met l'accent sur l'authentification forte (du client) pour assurer la sécurité d'une transaction commerciale électronique, à distance, et pour garantir la protection des utilisateurs.

La suite de ce chapitre va permettre de préciser les contours de l'utilisabilité, via les définitions mises en jeu dans les normes et les modèles. La synthèse des contributions issues des thèses que j'ai encadrées considérera l'utilisabilité à travers deux thèmes, à savoir l'authentification transparente sur mobile et la biométrie comportementale.

2 Normes et modèles

2.1 Normes ISO

L'utilisabilité est définie dans le standard ISO/IEC 9241-11:2018 *Ergonomie de l'interaction homme-système — Partie 11 : Utilisabilité — Définitions et concepts* de la manière suivante :

7. https://www.senat.fr/espace_presse/actualites/201803/directive_services_de_paiement_-dans_le_marche_interieur.html

Définition 4 (Utilisabilité)**.** *L'utilisabilité est le degré auquel un système, un produit ou un service peut être utilisé par des utilisateurs spécifiés pour réaliser des objectifs spécifiés avec efficacité, efficience et satisfaction dans un contexte d'utilisation spécifié.*

L'utilisabilité, en tant que résultat de l'interaction avec un système, un produit ou un service, repose donc sur trois propriétés :

- **efficacité** : précision et degré d'achèvement avec lesquels l'utilisateur atteint des objectifs spécifiés
- **efficience** : rapport entre les ressources utilisées (temps, effort, coûts, etc.) et les résultats obtenus
- **satisfaction** : degré selon lequel les réactions physiques, cognitives et émotionnelles de l'utilisateur qui résultent de l'utilisation d'un système, produit ou service répondent aux besoins et attentes de l'utilisateur

L'utilisabilité est pertinente lors de la conception ou de l'évaluation des interactions avec un système, un produit ou un service : les systèmes d'authentification biométrique sont donc particulièrement concernés par cette notion.
On trouve une déclinaison de ces trois propriétés au niveau de l'enrôlement. Il s'agit du rapport technique ISO/IEC Technical Report (TR) 29196 : 2018 *Guidance for Biometric Enrolment* [8], qui propose d'évaluer :

- la qualité de l'échantillon capturé pour mesurer l'efficacité ;
- le temps d'enrôlement et les erreurs pour mesurer l'efficience ;
- l'attitude de l'utilisateur, ses perceptions, son ressenti, son avis pour mesurer sa satisfaction

L'utilisabilité apparaît dans plusieurs normes détaillées dans le standard ISO/IEC JTC1 SC37 sur la biométrie, qui a déjà été mentionné au chapitre précédent. Cette notion est considérée comme un facteur clé dans la mise en œuvre des systèmes biométriques, même si aucune analyse standardisée n'existe à l'heure actuelle. Les travaux de normalisation ont lieu au sein des groupes de travail WG4 *Technical Implementation of Biometric Systems* et WG6 *Cross-Jurisdictional and Societal Aspects*. Ce dernier propose des lignes directrices dans le standard ISO/IEC TR 24714-1 : 2008 *Jurisdictional and societal considerations for commercial applications-Part 1 : General guidance* [9], selon trois aspects complémentaires :

- les questions juridiques en lien avec la vie privée de l'utilisateur et la protection des données personnelles ;
- l'accessibilité ;
- les questions de santé et de sûreté.

De façon plus pragmatique, le standard ISO/IEC 24779 [10] propose des icônes et des symboles dont la fonction est d'assister et guider les utilisateurs de systèmes biométriques, pour une meilleure interaction (présentation de la donnée biométrique, capture plus rapide, moins d'erreurs, etc.).

Dans l'article [Bla+19], les auteurs mentionnent le standard ISO/IEC 29156 : 2015 *Guidance for specifying performance requirements to meet security and usability needs in applications using biometrics* (établi par le JTC1 SC37). Y sont décrits des compromis nécessaires entre la sécurité

8. https://www.iso.org/obp/ui/fr/#iso:std:iso-iec:tr:29196:ed-2:v1:en
9. https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:24714:-1:ed-1:v1:en
10. https://www.iso.org/obp/ui/#iso:std:iso-iec:24779:-1:ed-1:v1:en

et l'utilisabilité dans les systèmes biométriques, en regard des autres mécanismes d'authentification, reposant sur des mots de passe ou des jetons. Dans ce même article, certains taux d'erreurs vus dans l'Introduction sont reliés aux concepts rattachés à l'utilisabilité. Par exemple, le FTE (*Failure To Enrol rate*) pourrait être mis en relation avec l'accessibilité et l'utilisabilité, le FTA (*Failure To Acquire rate*) avec les performances des capteurs et les facteurs humains et le FRR (*False Rejection Rate*) avec les questions d'ergonomie. Toujours dans cet article est évoqué le projet ISO/IEC 21472 *Scenario evaluation methodology for user interaction influence in biometric system performance*, au sein du groupe de travail ISO/IEC SC37 WG5, dont l'objectif est d'élaborer une méthodologie pour évaluer l'impact de trois facteurs sur les performances des systèmes biométriques : l'utilisateur, le système biométrique et leurs interactions. Le tableau 2.1, issu de l'article [Bla+19], répertorie toutes les normes ayant trait à l'utilisabilité et les systèmes biométriques.

Standard Identifier	Title
ISO 9241	Ergonomics of human-system interaction
ISO 25060	The common industrial format (CIF) for usability - General framework
ISO/IED 29196	Guidance for biometric enrolment
ISO/IEC PDTR 30125	Biometrics used with mobile devices
ISO/IEC 24714	Jurisdictional and societal considerations for commercial applications
ISO/IEC 24779	Information technology - Cross-jurisdictional and societal aspects of implementation of biometric technologies - Pictograms, icons and symbols for use with biometric systems
ISO/IEC 19794	Information technology - Biometric data interchange formats
ISO/IEC 29156	Guidance for specifying performance requirements to meet security and usability needs in application using biometrics
ISO/IEC 21472	Scenario evaluation methodology for user interaction influence in biometric system performance
ISO/IEC 19795	Biometric performance testing and reporting

Tableau 2.1 – Les différents standards sur les interactions avec un système biométrique [Bla+19]

Dans l'article [Bev+16], Bevan *et al.* soulignent que l'utilisabilité est pertinente dans les contextes suivants :

- une utilisation continue régulière, pour permettre aux utilisateurs d'atteindre leurs objectifs de manière efficace, efficiente et satisfaisante,
- l'apprentissage (mêmes objectifs pour les nouveaux utilisateurs),
- une utilisation peu fréquente (mêmes objectifs à chaque réutilisation),
- une utilisation par des personnes ayant les capacités les plus variées,
- une réduction maximale du risque et des conséquences indésirables des erreurs de l'utilisateur,
- la maintenance.

Le rapport technique de Speicher [Spe15] va plus loin et propose d'adapter les métriques de qualité issues d'un autre standard ISO/IEC 25010:2011 *Ingénierie des systèmes et du logiciel - Exigences de qualité et évaluation des systèmes et du logiciel - Modèles de qualité du système et du logiciel*, qui sont : (i) les métriques d'utilisabilité internes, (ii) les métriques d'utilisabilité externes et (iii) les métriques d'utilisabilité en usage. L'étude de l'utilisabilité d'un système biométrique relève de cette troisième famille de mesures.
La question sous-jacente à l'évaluation de l'utilisabilité peut être formulée ainsi : comment mesurer à quel point un système est utilisable ? La réponse à cette question se trouve dans le standard ISO/IEC TR 25060:2010 *Systems and software engineering — Systems and software product Quality Requirements and Evaluation (SQuaRE) — Common Industry Format (CIF) for*

usability : General framework for usability-related information. On y trouve des indications pour la spécification et l'évaluation de l'utilisabilité des systèmes interactifs (dans le sens des interactions avec les individus).
Dans la littérature, on trouve plusieurs modèles qui traitent de l'utilisabilité. J'ai choisi d'en présenter trois :

- le modèle de Nielsen, qui considère l'utilisabilité comme une composante de l'acceptabilité [Nie93] ;
- le modèle HBSI – *Human Biometric System Interaction* –, modèle conceptuel centré sur la relation entre l'utilisateur et le système biométrique [KED07] ;
- le modèle BioTAM – *Biometric Technology Acceptance Model* –, modèle prenant en compte les facteurs humains et sociaux qui jouent un rôle dans l'adoption des systèmes d'authentification biométrique, notamment la confiance [MPO13].

Ces trois modèles vont être présentés successivement, afin de mettre en valeur leur complémentarité.

2.2 Le modèle de Nielsen

Dans le livre [Nie93], Jakob Nielsen replace l'utilisabilité dans un contexte plus général, en tant que sous-partie de l'acceptabilité d'un système, comme illustré par la figure 2.1. Il s'agit d'un modèle global, qui ne concerne pas uniquement l'utilisabilité dans les systèmes biométriques.

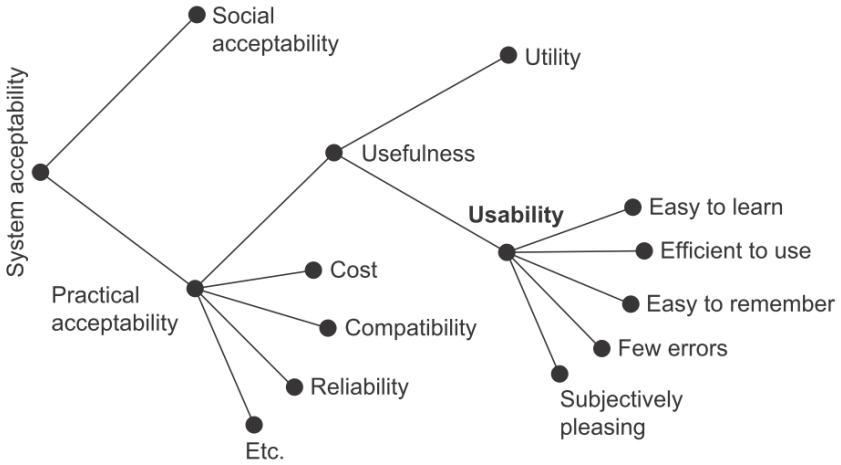


FIGURE 2.1 – Modèle de l'acceptabilité d'un système [Nie93]

Malgré son ancienneté – il a été publié en 1993 –, ce modèle fait toujours référence ; il établit que le concept d'utilisabilité repose sur cinq caractéristiques : la facilité d'apprentissage ; l'efficience, ou efficacité d'utilisation ; la facilité d'appropriation, de mémorisation ; la fiabilité (ou prévention des erreurs) ; la satisfaction.
Ces caractéristiques doivent être prises en compte dès la conception d'un système d'authentification biométrique.

1. Contrairement aux mots de passe et autres code PIN, la **facilité d'apprentissage** est une qualité intrinsèque de tout système biométrique ; c'est même un des arguments principaux pour le déploiement de telles solutions d'authentification.
2. L'**efficience** traduit la facilité avec laquelle l'utilisateur atteint son objectif, comme l'accès à un service, un site, un objet connecté. Ce critère repose sur plusieurs niveaux : la

qualité du capteur biométrique, la rapidité de l'algorithme de comparaison, la facilité de l'enrôlement, etc.

3. **La facilité d'appropriation, de mémorisation** est elle aussi inhérente aux systèmes biométriques : il suffit de poser son doigt sur le capteur, ou de présenter son visage devant la caméra, d'interagir avec un écran, etc. Il n'y a rien à mémoriser et aucun objet à présenter.
4. **La fiabilité** repose sur un faible taux d'erreurs : elle dépend donc totalement des performances du système biométrique. Ces performances reposent elles-mêmes sur la fiabilité du capteur biométrique, sur les performances de l'algorithme de reconnaissance.
5. **La satisfaction** quant à elle, traduit le ressenti de l'utilisateur. Elle dépend à la fois du profil de l'utilisateur (expérimenté ou non, habitué à une technologie particulière, etc.) et du contexte, des conditions d'utilisation du système d'authentification biométrique.

Dans un contexte d'usage, l'acceptabilité est synonyme de degré d'*intégration* et d'*appropriation* d'un objet, comme décrit par Barcenilla et Bastien dans l'article [BB09]. Ces deux propriétés paraissent essentielles lors de la conception, du développement puis du déploiement d'une solution d'authentification biométrique, afin de garantir une utilisation future importante. Elles sont définies comme suit :

- l'**intégration** correspond à la manière dont le produit, ou système technique, s'insère dans la chaîne instrumentale existante et dans les activités de l'utilisateur, et comment il contribue à transformer ces activités ;
- l'**appropriation** renvoie à la façon dont l'individu investit personnellement l'objet ou le système et dans quelle mesure celui-ci est en adéquation avec ses valeurs personnelles et culturelles, lui donnant envie d'agir sur ou avec celui-ci, et pas seulement de subir son usage. Le cas extrême de l'appropriation est celui où l'objet devient une composante de l'identité du sujet.

Tous ces critères contribuent à une expérience utilisateur (UX, *User eXperience*) optimale. Pour aller plus loin, l'*UX Design*, terme utilisé pour la première fois par Donald Norman dans les années 90, désigne toute expérience vécue en interaction avec un dispositif, numérique ou non. Le diagramme de la figure 2.2 indique que l'utilisabilité est bien l'une des composantes de l'*UX Design*.
Un principe énoncé par J. Nielsen : « *Designers are not users / People don't know what they need* » montre à quel point la confusion peut être facile entre utilisabilité et UX. Le lien entre les deux s'énonce de la manière suivante : le respect de différents critères (simplicité, lisibilité, utilisabilité, accessibilité, etc.), historiquement liés au développement web et à la conception d'interfaces utilisateurs (*UI Design*), contribuent à une expérience utilisateur positive.
L'utilisabilité devient importante dans la conception des solutions d'authentification biométrique car leurs usages se sont beaucoup diversifiés et intensifiés depuis quelques années (voir le paragraphe 2 de l'Introduction, page 5). Le modèle suivant replace la notion d'utilisabilité au cœur de l'usage des systèmes biométriques.

2.3 Le modèle HBSI : Human Biometric Sensor Interaction

Ce deuxième modèle, à la différence de celui de Nielsen – qui peut être appliqué à tout type de système –, concerne uniquement les systèmes d'authentification biométriques. Il a été développé à l'Université de Purdue et fait partie de la classe des HCI, ou *Human-Computer Interactions*. L'article de référence sur le sujet a été publié en 2007, par Kukula *et al.* [KED07]. Reconnaissant

FIGURE 2.2 – Une cartographie de l'UX Design par Daniel Würstl

que la biométrie peut sembler intrusive, l'article se concentre sur l'un des aspects-clés dans l'étude des interactions, à savoir le degré de confiance des individus dans la technologie. En effet, un manque de confiance peut compromettre toute chance de succès dans le déploiement d'un nouveau système et dériver vers une mauvaise utilisation, voire un rejet de la technologie. C'est un modèle conceptuel, représenté à la figure 2.3, centré sur la relation entre l'utilisateur et le système biométrique : l'interaction est cruciale, dans le sens où un utilisateur inexpérimenté peut avoir des difficultés à présenter sa donnée biométrique au capteur.

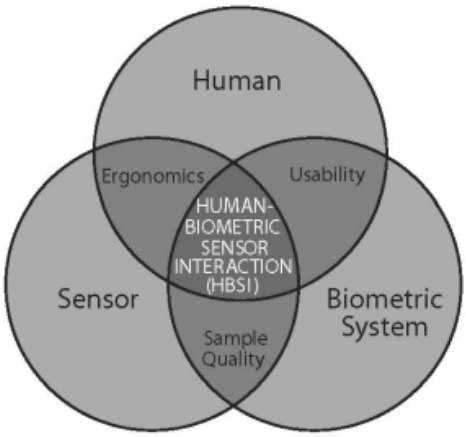


FIGURE 2.3 – Modèle HBSI [KED07]

Ce modèle s'intéresse à l'ergonomie, l'utilisabilité et le traitement du signal, la qualité des données, pour permettre de concilier la présentation des capteurs, du logiciel, la mise en œuvre des différents éléments du système d'authentification, etc., à l'utilisateur. Le modèle HBSI propose une méthode d'évaluation des performances du système à partir des technologies et des interactions, plus précisément il étudie l'impact des interactions avec l'utilisateur sur les performances globales du système, comme illustré à la figure 2.4. Si ces interactions ne sont pas prises en

compte, les performances optimales ne peuvent être atteintes en raison d'erreurs de type FTA (*Failure to Acquire*), FTE (*Failure to Enrol*), avec un réel impact sur le FRR.

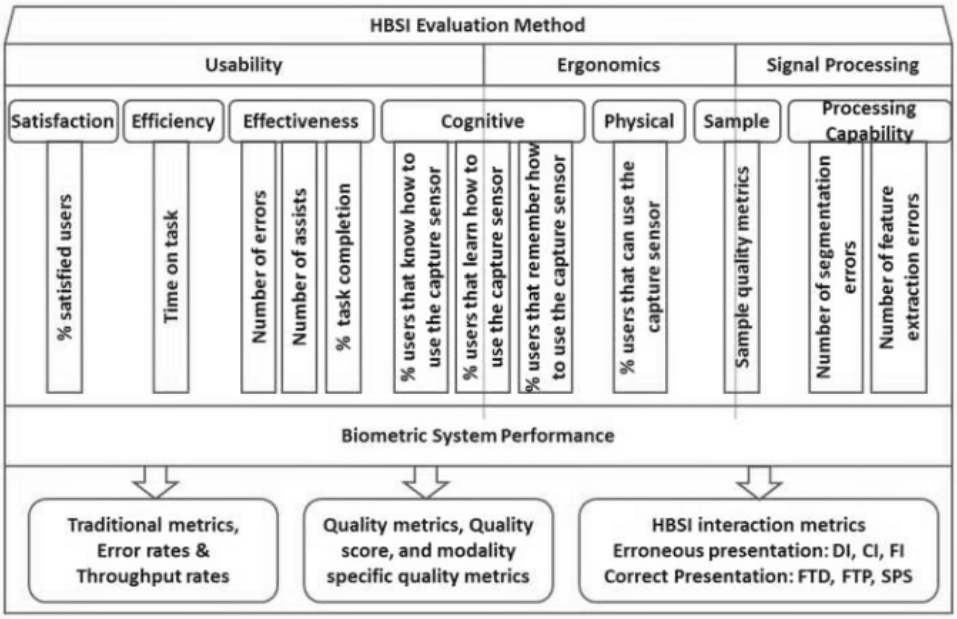


FIGURE 2.4 – Méthode d'évaluation HBSI [Mig+16]

La figure 2.5 offre un focus sur les cinq métriques d'erreurs liées aux interactions étudiées dans le framework HBSI. Ces erreurs sont définies dans par Brockly *et al.* dans l'article [Bro+09] :

- **DI = *Defective Interaction***
  Ce type d'erreur se produit lorsqu'une présentation incorrecte de la donnée n'est pas détectée par le système biométrique. Il s'agit donc d'une double erreur : de la part de l'utilisateur et du système. Le taux de DI permet de mesurer les erreurs des utilisateurs qui affectent le temps de passage devant le système et ses performances.
- **CI = *Concealed Interaction***
  Ce type d'erreur se produit lorsqu'une présentation incorrecte est détectée par le système biométrique, mais n'est pas classifiée correctement comme une erreur. Dans ce cas, une erreur est commise à la fois par l'utilisateur et par le système.
- **FI = *False Interaction***
  Ce type d'erreur se produit lorsqu'une présentation incorrecte est détectée par le système biométrique et, contrairement à une CI, est correctement traitée comme une erreur. L'étape suivante consiste à soit rejeter l'utilisateur, soit lui proposer une nouvelle tentative.
- **FTD = *Failure To Detect***
  Ce type d'erreur se produit lorsqu'une présentation correcte de sa donnée par l'utilisateur n'est pas détectée par le système. Même si la résultante est la même, la différence avec une interaction de type DI est que seul le système est en tort.
- **FTP = *Failure To Process***
  Ce type d'erreur se produit lorsque la présentation de la donnée par l'utilisateur est correcte, que le système la détecte, mais que les traitements suivants échouent dans la création d'un gabarit. Il peut s'agir d'un problème au niveau du traitement d'image, de l'extraction de caractéristiques, du contrôle de qualité, etc.

En supplément, le framework HBSI propose la métrique SPS (***Successfully Processed Sample***) : une présentation correcte de la donnée par l'utilisateur, traitée correctement par le système. Il s'agit du cas nominal où tout fonctionne correctement.

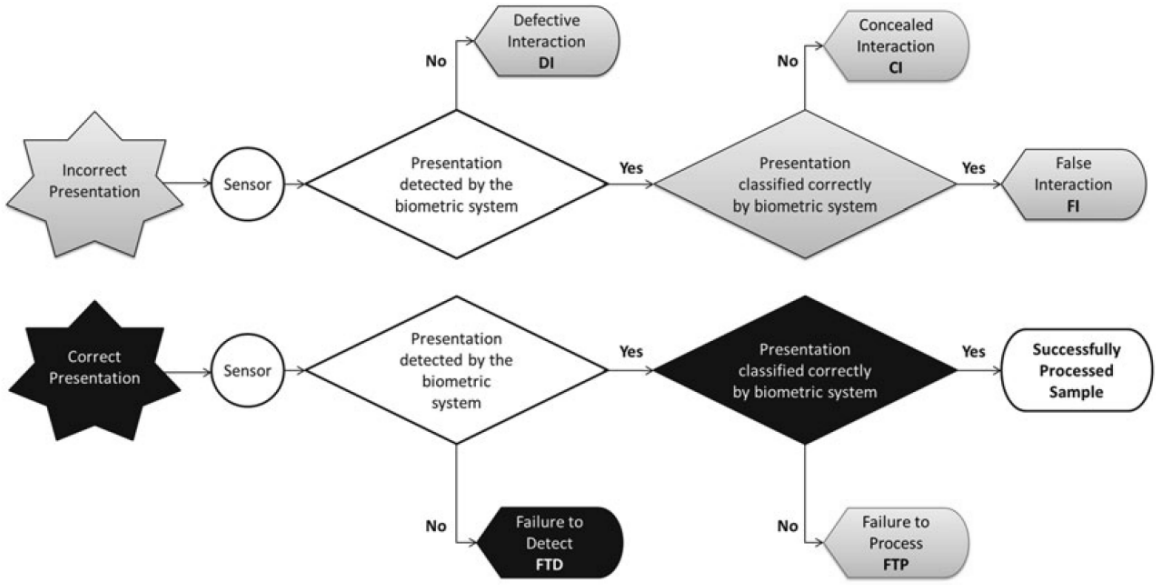


FIGURE 2.5 – Métriques d'erreurs du framework HBSI [Bro+09]

Le framework HBSI a été testé sur différentes modalités : l'empreinte digitale, la forme de la main, le visage, la dynamique de signature. Les références sont mentionnées dans l'article de Miguel-Hurtado *et al.* [Mig+16], qui étend la méthode à un système d'authentification sur smartphone reposant sur la reconnaissance de la voix et du visage. Les auteurs proposent un cadre pour l'évaluation des critères standard d'utilisabilité, définis par la norme ISO/IEC 9241-11 : 2018 (voir le début de ce chapitre) :

- *l'efficacité* est définie en fonction du temps passé à accomplir une tâche (enrôlement ou vérification), dès lors que les utilisateurs maîtrisent le processus ;
- *l'efficience* fait référence à (i) la mesure dans laquelle le système se comporte de la manière attendue par les utilisateurs et (ii) la facilité avec laquelle ils peuvent l'utiliser pour faire ce qu'ils veulent. Les mesures quantitatives reposent sur des indicateurs :
  - le pourcentage d'erreurs détectées ;
  - le pourcentage de recours à l'aide (par un opérateur) pendant la réalisation d'une tâche ;
  - le taux de succès de la première tentative pour chaque tâche ;
- *la satisfaction* est mesurée, de façon classique, à travers un questionnaire *a posteriori*

Pour les concepteurs de systèmes biométriques, le framework HBSI propose de combiner des aspects d'anthropométrie, d'ergonomie, de conception centrée sur l'utilisateur et d'utilisabilité. Voici quelques avantages qui peuvent être retirés du recours au modèle HBSI :

- diminuer la charge mentale de l'utilisateur,
- faciliter la mise en place d'habitudes, de bonnes pratiques,
- favoriser les interactions inconscientes, plus rapides et naturelles,
- placer le corps de l'utilisateur dans une position reproductible, sans stress,
- proposer des conditions de collecte optimales, une expérience utilisateur améliorée, afin

d'obtenir les meilleures performances.

L'article [Bla+19], déjà mentionné, présente un état de l'art sur ces sujets, depuis les origines du modèle HBSI, jusqu'aux nouvelles tendances. Il propose également un tour d'horizon des outils développés pour mesurer l'utilisabilité et l'acceptation des systèmes biométriques par les utilisateurs. Le recours à l'authentification biométrique pour combler les nouveaux besoins de sécurité doit englober la prise en compte de l'utilisabilité. Le troisième modèle, BioTAM, va encore plus loin et met en perspective les trois notions : sécurité, utilisabilité et respect de la vie privée.

2.4 BioTAM : *Biometric Technology Acceptance Model*

Si on replace les systèmes d'authentification biométriques parmi les SI (Systèmes d'Information), leur utilisabilité peut être mise en perspective par rapport à l'acceptation des technologies de l'information. Les grandes tendances dans ce domaine sont recensées par Miltgen *et al.* dans l'article [MPO13] : *Technology Acceptance Model* (TAM), *Diffusion Of Innovations* (DOI) et *Unified Theory of Acceptance and Use of Technology* (UTAUT). Ces différents modèles possèdent un point commun : ils suggèrent que, face à une nouvelle technologie, certains facteurs influencent la décision des utilisateurs quant à leurs usages. Parmi ces facteurs, on peut mentionner les croyances ou les ressentis à propos de ces technologies, la propension à les recommander, leur utilité perçue, leur utilisabilité, etc. C'est pourquoi cet article [MPO13] propose de combiner les trois approches les plus importantes pour étudier les rapports des utilisateurs avec les systèmes biométriques, c'est-à-dire les modèles TAM, DOI et UTAUT. Leur conclusion est assez surprenante : les facteurs qui influent le plus sur l'acceptation des systèmes biométriques et la capacité des utilisateurs à les recommander sont à rechercher du côté de la confiance et du respect de la vie privée. Kanak et Sogukpinar vont plus loin en incluant le facteur *confiance* dans un modèle comprenant également le compromis sécurité - respect de la vie privée et la disposition favorable des utilisateurs [KS17]. On retrouve des éléments des modèles standards, comme la facilité d'usage perçue, l'utilisabilité perçue, le ressenti vis-à-vis de la technologie, ou encore les intentions des utilisateurs en termes de comportement face à la solution d'authentification. La figure 2.6 illustre l'approche proposée dans cet article.

Le BioTAM repose sur un ensemble d'hypothèses, détaillées à la figure 2.7, en lien avec la figure 2.6. Le but est d'évaluer l'utilisation réelle du système (notée *A : Actual System Use*), grâce à différents éléments liés aux systèmes biométriques.

- **L'utilité perçue, *Perceived usefulness (U)***
  Elle est améliorée par des services personnalisés, une communication efficace, des mécanismes d'avertissement, etc. La confiance, *Trust (T)*, a également un impact direct sur *U*, ainsi que les performances en terme de reconnaissance de l'utilisateur.
- **La facilité d'usage perçue, *Perceived ease of use (E)***
  Elle est liée à une absence d'effort dans les différentes étapes des systèmes biométriques. On touche ici à l'utilisabilité, car *E* repose sur des variables telles que les menus, les icônes, les écrans, les capteurs conviviaux, etc., en résumé, tout ce qui touche à l'*UI Design*. Elle est influencée (négativement) par les erreurs d'acquisition et d'enrôlement. Si les utilisateurs ont l'assurance d'une grande facilité d'usage, leur inclination à poursuivre l'utilisation (*Public willingness*) du système augmente.
- **Les facteurs externes, *External factors***
  Il revient à chaque concepteur de système biométrique de définir un ensemble de facteurs externes influençant *U* et *E*. Par exemple, l'aisance vis-à-vis des nouvelles technologies a tendance à diminuer lorsque l'âge des utilisateurs augmente. Ou encore d'autres facteurs sociaux ou humains, comme le niveau d'études, les conditions de travail (qui peuvent dé-

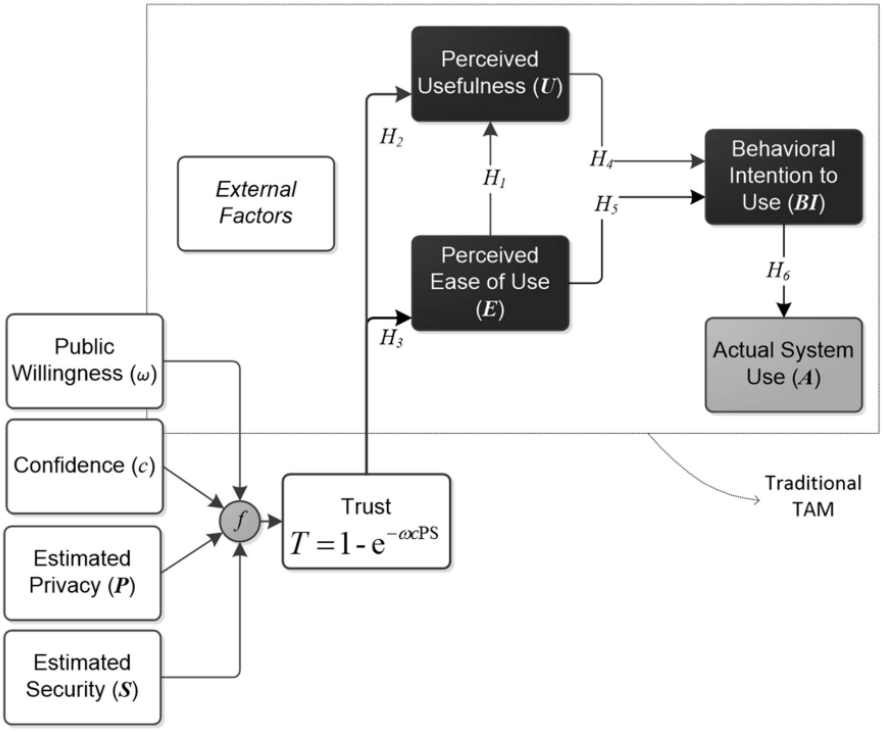


FIGURE 2.6 – Biometric Technology Acceptance Model, extrait de [KS17]

tériorer les empreintes digitales), la présence d'un service d'assistance de qualité 24h/24, etc.

Ces éléments font partie des TAM (*Technology Acceptance Models*) traditionnels, comme cela est représenté à la figure 2.6. La nouveauté proposée par les auteurs de l'article [KS17] se situe en bas à gauche de cette figure. Il s'agit de prendre en compte l'influence de la confiance (*Trust,* T) sous la forme d'une fonction dépendant de la volonté du public (*Public willingness* ω), de l'assurance (*Confidence* c), du respect de la vie privée estimé (*Estimated privacy* P) et du niveau de sécurité estimé (*Estimated security* S) :

$$T = 1 - e^{-\omega cPS} \tag{2.1}$$

Plus précisément, T reflète le niveau de confiance des utilisateurs dans le système déployé, en fonction de la croyance des utilisateurs en ce qui concerne le niveau de sécurité et de respect de la vie privée. La fonction (2.1) a été proposée pour répondre à un certain nombre de propriétés désirées et les variations des variables sont étudiées en détails d'un point de vue mathématique dans l'article [KS17]. En pratique, les valeurs de ces variables sont déduites à partir des réponses des utilisateurs à un questionnaire précis. Les auteurs discutent également de la distinction entre les notions de *trust* et de *confidence*, toutes deux traduites par le terme *confiance* en français. Ils s'appuient en partie sur l'approche développée par Miltgen *et al.* [MPO13] :
« *Trust in the technology does not only influence the perceived risks associated with accepting the technology but is also an antecedent of ease of use and usefulness.* »
Comme indiqué plus haut, le tableau 2.7 détaille les hypothèses proposées H_1 à H_6 pour mettre en relation les facteurs qui influencent l'usage réel du système biométrique par les utilisateurs.

H_1	perceived ease of use (*E*) positively influences the perceived usefulness (*U*)
H_2	high trust (*T*) on a BAS will lead to increased perceived usefulness (*U*)
H_3	high trust (*T*) on a BAS will lead to increased perceived ease of use (*E*)
H_4	perceived usefulness (*U*) while using a BAS has a positive effect on users' behavioural intention (*BI*) to use the system
H_5	perceived ease of use (*E*) while using a BAS has a positive effect on users' behavioural intention (*BI*) to use the system
H_6	behavioural intention (*BI*) of users to use a BAS positively influences the actual usage (*A*)

FIGURE 2.7 – Hypothèses du BioTAM, extrait de [KS17]. BAS = *Biometric Authentication System*

Les articles cités précédemment insistent sur le fait que l'utilisation effective d'un système biométrique est influencée par des facteurs d'ordre pratique ou psychologique, comme les interfaces utilisateurs, les procédures d'enrôlement et de vérification, les appareils, et d'autres outils auxiliaires. Les auteurs ne mentionnent pas toujours explicitement la notion d'utilisabilité (on trouve plutôt l'expression *user acceptance*), mais les facteurs précédents ont un lien évident avec cette caractéristique essentielle d'un système biométrique.

3 Discussion

On constate que l'utilisabilité est une propriété indispensable – mais non suffisante – à la réussite du déploiement d'un système d'authentification biométrique. Le ressenti de l'utilisateur en matière de sécurité, de respect de la vie privée et d'utilisabilité conditionne l'adoption de la technologie, et même la confiance dans la technologie. C'est pourquoi l'étape de conception ne doit pas faire l'impasse sur l'étude de l'utilisabilité du système à développer. Un point de vue plus large est développé dans la thèse de Maureen Sullivan [Sul12], qui reprend le modèle UTAUT (*Unified Theory of Acceptance and Use of Technology*), illustré à la figure 2.8.
Ce modèle met en évidence l'influence de certains paramètres sur le comportement final de l'utilisateur : les attentes en matière de performance, en matière d'effort à produire, l'influence sociale (l'utilisateur prend en compte l'avis de personnes importantes à ses yeux par rapport à la technologie proposée), les conditions favorables (à l'usage de la technologie). On peut donc constater que des facteurs sociaux, sociétaux et psychologiques devraient être pris en compte. Cette tendance commence à émerger réellement, comme on peut le voir dans un rapport publié par Mastercard en 2020 [11]. Ce rapport reprend les conclusions du modèle HBSI de l'Université de Purdue, et présente une critique constructive des modalités étudiées dans les articles cités dans le paragraphe 2.3, à savoir l'empreinte digitale, le visage, la forme de la main et la voix. Voici une citation extraite de ce rapport, page 7 :
« *No matter how accurate biometrics are, they will only be effective if people trust them and want to use them.* »
La prise en compte de l'utilisabilité est donc en progression, sous l'influence des plus grands acteurs industriels capables de révolutionner les usages de l'authentification biométrique.

La dernière partie de ce chapitre est consacrée à la présentation de certaines contributions ayant trait à l'utilisabilité des systèmes d'authentification biométrique.

11. https://www.mastercard.com/news/research-reports/2020/evolution-of-biometrics/

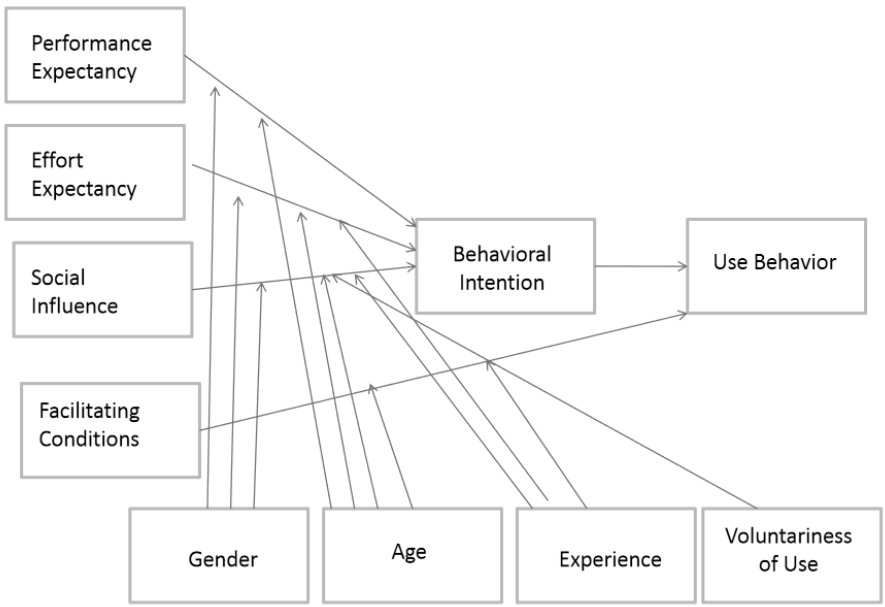


FIGURE 2.8 – Modèle UTAUT

4 Contributions

Parmi les trois propriétés des systèmes biométriques qui structurent ce manuscrit, l'utilisabilité est celle que j'ai le moins explorée à part entière. Elle est cependant présente dans de nombreux aspects de mes recherches. Cette synthèse de mes contributions liées à l'utilisabilité va prendre la forme d'une succession de pistes dans lesquelles elle est apparue, notamment au fil des thèses encadrées. Même si cette partie est moins consistante que dans les deux autres chapitres, cette thématique a progressé en arrière-plan dans mes réflexions, pour trouver une place centrale dans mon projet de recherche. Il me semblait donc impossible d'occulter ces points d'accroche, d'ancrage, ces éclairages ponctuels qui, de détails, sont devenus des incontournables.
Les deux principales thématiques au sein desquelles l'utilisabilité a émergé sont : l'authentification continue et transparente sur mobile et la biométrie comportementale, plus précisément la dynamique de frappe au clavier.

4.1 L'utilisabilité et l'authentification continue sur mobile

L'utilisabilité du couple traditionnel {login + mot de passe} décroît de plus en plus. En cause, les politiques toujours plus strictes de définition de mots de passe forts : l'expansion des services numériques fait exploser le nombre de mots de passe par utilisateur, et ces mots de passe sont toujours plus complexes. Le fardeau est de moins en moins supportable pour la mémoire des utilisateurs, qui adoptent des comportements non souhaitables du point de vue de la sécurité : le même mot de passe est utilisé pour plusieurs sites, les mots de passe choisis sont simples (si possible), écrits sur des feuilles qui traînent près de l'ordinateur, dans le portefeuille, etc. A défaut de prise en compte de ses besoins, de ses attentes, – et encore moins de son fonctionnement cognitif –, l'utilisateur recrée un semblant d'utilisabilité avec ces stratégies qui lui sont propres. Il s'agit de véritables stratégies de contournement qui lui permettent de compenser, de s'adapter à :

- un manque criant d'homogénéité entre les systèmes d'authentification,
- des enjeux de sécurité pas très clairs,

- trop de contraintes (des majuscules, des minuscules, des caractères spéciaux, une longueur minimale, des mots de passe différents... à définir et à retenir par dizaines).

La conception est trop peu souvent centrée sur l'utilisateur. Pourtant, comme on l'a vu dans le début de ce chapitre, c'est une des clés du succès pour le déploiement de toute nouvelle technologie. D'un autre côté, les cyberattaques sont de plus en plus fréquentes. Aussi, d'autres solutions d'authentification sont plébiscitées, notamment celles reposant sur la biométrie. Les systèmes biométriques embarqués sur smartphone depuis quelques années ont relégué le code PIN historique au rang de solution de secours, lorsque la reconnaissance d'empreinte digitale ou de visage ne fonctionne pas. L'étape suivante, intégrée aujourd'hui dans la version 2 du protocole 3D-secure, consiste à capturer un maximum de données pendant l'utilisation du smartphone, pour, d'une part, créer un modèle du comportement et des habitudes de l'utilisateur, et pour, d'autre part, établir un score de confiance dans l'identité de l'utilisateur, et ce, de façon transparente, sans le déranger.

Contexte

La thèse de Julien Hatin [Hat17] se situe dans ce contexte de l'authentification transparente continue sur mobile et propose des solutions respectueuses de la vie privée. Elle a pour titre *Evaluation de la confiance dans un processus d'authentification*. Julien a soutenu sa thèse en 2017. Il s'agit d'une thèse CIFRE, dirigée par Christophe Rosenberger côté GREYC et par Jean-Jacques Schwartzmann côté Orange Labs, que j'ai co-encadrée. Le point de départ des travaux de Julien Hatin est le constat suivant : plutôt que de demander à l'utilisateur de mémoriser une donnée, tâche difficile pour un individu et simple pour un ordinateur, pourquoi ne pas demander à une machine d'apprendre à reconnaître en continu son utilisateur ? Toutefois, ce changement de paradigme impose de respecter un certain nombre de règles. Tout d'abord, ces authentifications ne doivent pas constituer une intrusion dans la vie privée de l'utilisateur : la garantie de transparence doit être maximale. Une authentification en continu ne doit pas se transformer en espionnage, au risque de briser la confiance de l'utilisateur dans les services qui lui sont proposés. Il est donc nécessaire de redéfinir les mécanismes d'authentification pour permettre cette connexion permanente. Cela passe par une authentification continue de l'utilisation du smartphone pour un meilleur confort d'usage et une meilleure sécurité.
Julien Hatin a proposé dans sa thèse trois solutions d'authentification continue, transparente et respectueuse de la vie privée de l'utilisateur. Il a tenté de concilier les trois propriétés {sécurité + utilisabilité + vie privée}. Ces solutions reposent sur des techniques différentes qui garantissent de façon inhérente la protection des données personnelles récoltées sur le smartphone de l'utilisateur. La première solution utilise un histogramme des habitudes des utilisateurs masquées à l'aide de fonctions de hachage. La deuxième méthode utilise du chiffrement homomorphe pour protéger les données privées de l'utilisateur. La dernière méthode utilise le BioHashing, algorithme de biométrie révocable qui sera présenté au chapitre suivant. Pour des raisons d'utilisabilité, je vais uniquement présenter le principe de cette troisième approche.

Données

Cette méthode propose une nouvelle architecture client-serveur. Le smartphone joue le rôle du client et collecte les données comportementales de l'utilisateur quand celui-ci passe un appel ou envoie un SMS. Les données collectées peuvent être choisies parmi les suivantes : durée d'appel, position du téléphone (données de l'accéléromètre, du gyroscope), localisation du téléphone, le numéro du correspondant, l'environnement sonore de l'utilisateur.

Grâce au contexte favorable de la convention CIFRE, Julien a eu un accès (restreint) à des comptes-rendus d'appels (CRA) sécurisés, anonymisés par son entreprise. La base de données utilisée pour évaluer ces travaux contient l'historique des communications de cent personnes pendant un mois, notamment les informations suivantes :

- la latitude et la longitude de l'antenne,
- le numéro de l'appelant,
- le numéro de l'appelé,
- le type (appel ou SMS).

Au final, pour tenir compte des bases publiques dans ce domaine, les données qui ont été effectivement exploitées dans la thèse sont :

- la position géographique de l'antenne correspondant à l'appel
- le numéro de téléphone de la personne appelée

Les données de test proviennent donc soit de la base de Julien, soit d'une base publique du MIT de 2004, utilisée dans l'article de Li *et al.* [Li+11], collectée sur un modèle de téléphone mobile – on ne parle encore pas de smartphone – bien plus ancien. L'utilisation d'une base publique permet une comparaison à l'état de l'art. Dans l'approche avec protection des données par le BioHashing, seule la base CRA a été testée.

Architecture

Les données collectées sont protégées par l'algorithme de BioHashing avant d'être transmises via un canal sécurisé (connexion TLS par exemple), comme illustré à la figure 2.9. De façon très schématique, le BioHashing génère une donnée transformée, appelée BioCode, à partir d'une donnée de l'utilisateur et d'une clé secrète. On verra au chapitre suivant que l'algorithme de BioHashing a été conçu pour protéger des données biométriques. Dans le cadre de la thèse de Julien, il a été appliqué à des données de contexte.Pour garantir la protection des données personnelles, le secret est stocké dans l'élément sécurisé du téléphone.

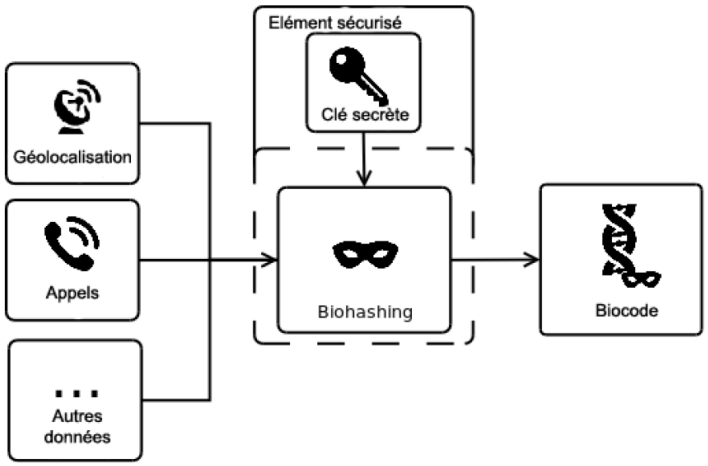


FIGURE 2.9 – Architecture du côté client, extrait de [Hat17]

Côté serveur, les BioCodes sont reçus en continu : un BioCode est envoyé par le client à chaque appel. Comme illustré à la figure 2.10, l'enrôlement est réalisé via la création d'un *template* à partir d'un nombre suffisant de BioCodes. Ce *template* est ensuite stocké dans une base de données. Ensuite le serveur peut passer en mode vérification, réalisée à l'aide d'un classifieur, de type SVM, ou de type K plus proches voisins (*K Nearest Neighbours*, KNN). Les détails de

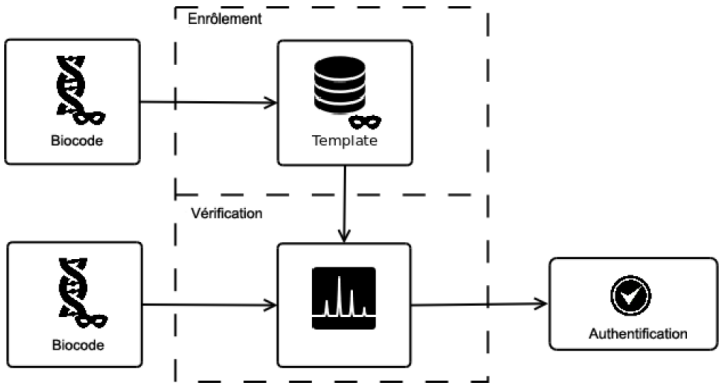


FIGURE 2.10 – Architecture du côté serveur, extrait de [Hat17]

l'expérimentation sont disponibles dans la thèse de Julien [Hat17] et dans l'article [Hat+17] (la mise en forme des données notamment, les bases de données étudiées, etc.). Parmi les tests effectués, deux scénarios ont été considérés : le scénario idéal, sans attaque, appelé « *meilleur des cas* » ; et le scénario avec attaque (vol du téléphone, donc l'attaquant a accès à la clé de l'utilisateur), appelé « *pire des cas* ». Dans le scénario sans attaque, seul l'algorithme des K plus proches voisins donne de bons résultats, présentés dans le tableau 2.2.

Nombre de voisins	EER (%)	seuil correspondant
1	1.04	0.30
2	1.10	0.62
3	1.09	0.95
4	1.16	1.29
5	1.19	1.63

Tableau 2.2 – Résultats du KNN dans le meilleur des cas

Dans le scénario avec attaque (vol du secret), l'algorithme des K plus proches voisins donne encore une fois les meilleurs résultats, présentés dans le tableau 2.3.

Nombre de voisins	EER (%)	Seuil correspondant
1	10.45	0.23
2	10.16	0.47
3	10.65	0.72
4	10.69	0.98
5	10.76	1.24

Tableau 2.3 – Résultats du KNN dans le pire des cas

On constate que l'EER a quasiment été multiplié par dix. C'est assez classique d'observer une telle dégradation des performances du BioHashing lors d'un vol du secret : on retrouve les performances du système d'authentification sans protection.

Perspectives

Ce qu'on peut retenir de cette méthode développée dans la thèse de Julien Hatin, c'est la simplicité de mise en œvre de l'architecture proposée. Les données sont collectées de façon transparente pour l'utilisateur, puis protégées par l'algorithme de BioHashing. Celui-ci présente également l'avantage d'être peu gourmand en quantité et en temps de calculs, ce qui favorise indéniablement l'utilisabilité, en comparaison des deux autres méthodes proposées dans la thèse. Une comparaison avec les résultats de l'état de l'art est reportée dans un tableau à la page 85 du manuscrit [Hat17]. Pour aller plus loin, Julien Hatin a proposé un mécanisme complet d'évaluation du niveau d'authentification, c'est-à-dire d'évaluation de la confiance dans l'identité numérique de l'utilisateur. Ce mécanisme est totalement adapté au cadre de l'authentification continue, et reste donc transparent pour l'utilisateur, en accord avec une bonne utilisabilité. Il repose sur la théorie de l'évidence de Dempster-Schaffer, et permet de combiner différentes preuves – comportant une part plus ou moins grande d'incertitude – apportées par la collecte de données en continu sur le smartphone de l'utilisateur. Les perspectives naturelles de ces travaux se situent au niveau de l'exploitation de nouvelles données comportementales, biométriques et donc plus représentatives de l'utilisateur que des habitudes d'appel. Il serait intéressant de tester les algorithmes et architectures proposés par Julien Hatin pour la protection de telles données. L'étape suivante pourrait être d'étudier la robustesse de ces schémas à d'autres attaques, au-delà du simple vol de clé.

4.2 Biométrie comportementale et utilisabilité

Il est indéniable que les modalités de biométrie comportementales présentent une meilleure utilisabilité que les modalités physiques ou physiologiques. En effet, la reconnaissance du rythme de frappe sur un clavier, de la démarche, ou encore de la façon de signer, les interactions avec un écran tactile, etc., semblent bien moins intrusives que la reconnaissance du visage ou d'une empreinte digitale, sans parler de l'ADN. Ce qui soulève des objections en matière de données comportementales, c'est plutôt la collecte elle-même : le plus souvent, elle est réalisée via un smartphone, capable de capturer sans discernement un flux de données personnelles, sans que l'utilisateur en soit informé ni conscient potentiellement. Heureusement, la réglementation évolue, et grâce au RGPD notamment, les acteurs du numérique sont censés indiquer les finalités de la collecte de données, et appliquer des principes de proportionnalité et de minimisation. Pour l'instant, je laisse volontairement de côté ces liens entre les données comportementales et le respect de la vie privée. Ils seront largement traités dans le chapitre suivant, et sont à l'origine de mon projet de recherche, détaillé dans le chapitre 4.

Le défi de la biométrie comportementale : réconcilier variabilité et utilisabilité

La biométrie comportementale met en jeu des aspects *acquis* de notre personnalité, qui continuent à évoluer au fil du temps. Cette évolution a lieu selon différentes temporalités et peut prendre plusieurs formes.

- **Une variabilité intra-utilisateur**
  Cette variabilité touche n'importe quelle modalité biométrique, pour différentes raisons :
  - une variation dans la présentation de la donnée au capteur (par exemple l'empreinte digitale),
  - une variation de la donnée elle-même (visage avec des maquillages différents, des lunettes différentes, une moustache ou une barbe qui pousse)

  S'agissant des modalités comportementales, elles présentent des causes supplémentaires de variations :

    - l'état de fatigue qui évolue au cours de la journée,
    - l'usage d'une main différente, ou d'une seule main, à cause d'une blessure (pour la DDF),
    - l'état d'esprit, les émotions, l'humeur et d'autres aspects psychologiques ou cognitifs.
- **Un vieillissement du modèle de référence**
  Dans le cas de certaines modalités physiques, comme le visage bien sûr, ou l'empreinte digitale, on constate un vieillissement de la donnée elle-même. S'agissant des modalités comportementales, on parle de vieillissement du modèle stocké comme référence, pour désigner l'écart qui se creuse au fil du temps entre le comportement capturé à un instant initial et le comportement actuel de l'utilisateur. Par exemple, la frappe au clavier a plutôt tendance à s'accélérer lorsque l'utilisateur passe du temps sur son ordinateur. Cette dérive par rapport au comportement initial peut être compensée par des mécanismes de mise à jour ou adaptation du modèle (*cf.* le paragraphe 3.2 du chapitre 1). Contrairement à la variabilité intra-utilisateur, cette variabilité nécessite un temps plus long pour apparaître.
- **Un apprentissage tout au long de la vie**
  Cette variabilité rejoint un peu la précédente. Elle comporte en plus une notion de maîtrise de l'outil de collecte, donc des aspects cognitifs supplémentaires. Cet apprentissage peut poser des problèmes de validité du modèle de référence, mais il permet également de capturer un comportement plus stable, car maîtrisé, de la part de l'utilisateur.

La biométrie comportementale paie en quelque sorte son acceptabilité accrue par une variabilité plus ou moins importante, qui affecte inévitablement ses performances, comparées à celles des modalités physiques ou physiologiques. La collecte elle-même des données comportementales, sur ordinateur ou smartphone le plus souvent, présente une grande utilisabilité intrinsèque, pour plusieurs raisons.

1. Il s'agit de continuer à faire des actions que l'on fait déjà : l'utilisateur va donc avoir confiance *a priori* dans ses compétences à interagir avec le système.

2. Le fait que la collecte soit réalisée sur des objets qui appartiennent à l'utilisateur (ou identiques aux siens si l'enrôlement se fait auprès d'un opérateur spécifique) renforce d'autant plus sa confiance.

3. Le stockage des données dans un élément sécurisé donne un sentiment de confiance à l'utilisateur, contrairement aux bases de données centralisées.

La thèse d'Abir Mhenni, largement évoquée au paragraphe 3.2 du chapitre 1, a proposé, entre autres contributions, un système adaptatif d'authentification biométrique par dynamique de frappe au clavier dont la phase d'enrôlement ne comporte qu'un seul échantillon (ou deux, dans le schéma plus complet de la mise à jour personnalisée grâce au zoo de Doddington). L'utilisateur n'est donc pas obligé de taper son mot de passe cinq, dix, ni vingt fois comme dans la quasi-totalité des schémas de la littérature. On peut donc parler d'authentification transparente par dynamique de frappe : on se ramène vraiment, au niveau de la charge cognitive, au cas où l'utilisateur définit son mot de passe, et éventuellement le tape une seconde fois pour confirmation. Pour pallier le manque d'information par rapport à un nombre de saisies plus important, la technique mise au point par Abir consiste à accumuler les saisies suivantes dans un template agglomérant dix échantillons au fil du temps, par un processus additif, de *growing window*. La stratégie de mise à jour du modèle permet ensuite de compenser les variabilités intrinsèques de la dynamique de frappe au clavier, comme on l'a vu précédemment.

Perspectives

La biométrie comportementale ne remplace pas un mot de passe ni d'autre facteurs d'authentification. Elle permet néanmoins de réduire la charge qui leur incombe de protéger les données sensibles. La sécurité d'un mot de passe – aussi fort soit-il –, ne repose que sur le secret. En offrant une couche supplémentaire et continue de confiance dans l'identité de l'utilisateur, la biométrie comportementale empêche le mot de passe d'être un point unique de défaillance de la sécurité. Plus largement, les smartphones actuels permettent de collecter de nombreuses interactions entre l'utilisateur et l'écran tactile ou les différents capteurs embarqués. Les perspectives pourraient consister à définir un mode de collecte de certaines données comportementales, de façon minimale, avec extraction du minimum d'information nécessaire pour authentifier le propriétaire du smartphone. Ceci afin de garantir le respect de sa vie privée. Cette perspective sera largement abordée dans mon projet de recherche.

5 Conclusion

Ainsi, même si la sécurité et le respect de la vie privée sont désormais indissociables de toute solution d'authentification biométrique, l'utilisabilité représente une troisième composante incontournable à prendre en compte dès la conception d'un tel système : les aspects sociaux, psychologiques, sociétaux, d'ergonomie, etc., ne peuvent être ignorés et contribuent largement à l'adoption du système par les utilisateurs. Un juste équilibre entre la sécurité et l'utilisabilité favorise la confiance de ceux-ci et améliore l'expérience utilisateur.
La conception centrée sur l'expérience utilisateur, ou *UX-design*, débouche sur la recherche centrée sur l'expérience utilisateur, ou *UX-research*. Voici une définition de l'*UX-research* [12] : « elle regroupe les différentes méthodes qui visent à étudier les usages et les besoins des utilisateurs ». L'*UX researcher* utilise de nombreuses méthodes — observation, entretiens, tests — pour comprendre, évaluer et améliorer l'expérience utilisateur des produits et services. On retrouve sur ce site une cartographie de l'UX, qui détaille cinq grandes phases : découverte, concept, organisation, design, production. La conception d'une nouvelle solution d'authentification biométrique ne peut plus faire l'économie de l'UX design/research.
Pour aller plus loin, on peut mentionner la norme ISO 9241-210 :2019 *Ergonomie de l'interaction homme-système - Partie 210 : Conception centrée sur l'opérateur humain pour les systèmes interactifs* qui énonce six principes :

- la conception est fondée sur une compréhension explicite des utilisateurs, tâches et environnements ;
- les utilisateurs sont impliqués tout au long de la conception et du développement ;
- la conception est pilotée et redéfinie par une évaluation centrée utilisateur ;
- le processus est itératif ;
- la conception s'adresse à toute l'expérience utilisateur ;
- l'équipe de conception inclut des compétences pluridisciplinaires

On retrouve dans ces préconisations des liens avec les méthodes Agiles, que j'ai commencé à expérimenter dans deux Projets 2A à l'Ensicaen. A la suite d'un des projets, j'ai proposé un stage à deux étudiants de l'Ensicaen, en lien avec mon projet de recherche : leurs sujets portent sur le développement d'un outil de collecte de données comportementales, respectueux de la vie privée. Au programme : de la méthodologie Agile (je vais travailler avec deux étudiants, un ingénieur de recherche, un ancien post-doctorant turc devenu maître de conférences), de l'utilisabilité et de la recherche plus fondamentale en biométrie (sécurité et vie privée).

12. https://www.usabilis.com/definition-ux-research/#uxresearch

On a vu précédemment que les nouveaux usages de la biométrie proviennent des différentes réglementations et de l'essor des smartphones et autres objets connectés. La relation entre les utilisateurs et les systèmes biométriques est par conséquent en perpétuelle évolution. On constate que les utilisateurs adoptent l'authentification biométrique lorsqu'elle facilite l'usage d'applications, de nouvelles technologies, tout en restant assez critiques vis-à-vis de la sécurité et de l'utilisabilité elle-même. Il faut absolument ajouter à ces deux critères le respect de la vie privée et la protection des données personnelles, qui font l'objet du chapitre suivant.

Références du chapitre 3

[BB09] J. BARCENILLA et J. BASTIEN. « L'acceptabilité des nouvelles technologies : quelles relations avec l'ergonomie, l'utilisabilité et l'expérience utilisateur ? » In : *Le travail humain* vol 72 (2009).

[Bev+16] N. BEVAN, J. CARTER, J. EARTHY, T. GEIS et S. HARKER. « New ISO Standards for Usability, Usability Reports and Usability Measures ». In : *Proceedings, Part I, of the 18th International Conference on Human-Computer Interaction. Theory, Design, Development and Practice*. 2016, p. 268-278.

[Bla+19] R. BLANCO-GONZALO, O. MIGUEL-HURTADO, C. LUNERTI, R. GUEST, B. CORSETTI, E. ELLAVARASON et R. SANCHEZ-REILLO. « Biometric Systems Interaction Assessment: The State of the Art ». In : *IEEE Transactions on Human-Machine Systems* (2019).

[Bla+13] R. BLANCO-GONZALO, R. SANCHEZ-REILLO, O. MIGUEL-HURTADO et J. LIU-JIMENEZ. « Usability analysis of dynamic signature verification in mobile environments ». In : *International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2013.

[Bro+09] M. BROCKLY, S. ELLIOTT, R. GUEST et R. B. GONZALO. « Human-Biometric Sensor Interaction ». In : *Encyclopedia of Biometrics*. Sous la dir. de S. Z. LI et A. K. JAIN. Springer US, 2009, p. 1-10.

[Cap20] E. A. CAPRIOLI. « L'identité numérique. Quelle définition pour quelle protection ? » In : sous la dir. de SOUS LA DIRECTION DE JESSICA EYNARD. LARCIER, 2020. Chap. Partie 2 - Identité numérique : les points d'ombre du règlement eIDAS, p. 123-138.

[Hat17] J. HATIN. « Évaluation de la confiance dans un processus d'authentification ». Thèse de doct. École doctorale mathématiques, information et ingénierie des systèmes, Caen, 2017.

[Hat+17] J. HATIN, E. CHERRIER, J.-J. SCHWARTZMANN et C. ROSENBERGER. « Privacy Preserving Transparent Mobile Authentication ». In : *International Conference on Information Systems Security and Privacy (ICISSP)*. Porto, Portugal, fév. 2017.

[KS17] A. KANAK et I. SOGUKPINAR. « BioTAM: A Technology Acceptance Model for Biometric Authentication Systems ». In : *IET Biometrics* 6 (2017).

[KSE10] E. P. KUKULA, M. J. SUTTON et S. J. ELLIOTT. « The human-biometric-sensor interaction evaluation method: Biometric performance and usability measurements ». In : *IEEE Transactions on Instrumentation and Measurement* (2010).

[KED07] E. KUKULA, S. ELLIOTT et V. DUFFY. « The Effects of Human Interaction on Biometric System Performance ». In : *First International Conference on Digital Human Modeling (ICDHM)*. 2007.

[Li+11] F. LI, N. CLARKE, M. PAPADAKI et P. HASKELL-DOWLAND. « Behaviour Profiling for Transparent Authentication for Mobile Devices ». In : *European Conference on Information Warfare and Security (ECIW)*. 2011.

[Mig+16] O. MIGUEL-HURTADO, R. BLANCO-GONZALO, R. GUEST et C. LUNERTI. « Interaction evaluation of a mobile voice authentication system ». In : *IEEE International Carnahan Conference on Security Technology (ICCST)*. 2016.

[MPO13] C. MILTGEN, A. POPOVIČ et T. OLIVEIRA. « Determinants of end-user acceptance of biometrics: Integrating the "Big 3" of technology acceptance with privacy context ». In : *Decision Support Systems* (2013).

[Nie93] J. NIELSEN. *Usability Engineering*. Morgan Kaufmann Publishers Inc., 1993.

[Spe15] M. SPEICHER. *What is Usability? A Characterization based on ISO 9241-11 and ISO/IEC 25010*. Rapp. tech. 2015.

[Sul12] M. S. SULLIVAN. « A study of the relationship between personnality types and the acceptance of Technical Knowledge Management Systems (TKMS) ». Thèse de doct. Capella University, 2012.

CHAPITRE

3

Biométrie et vie privée

If this is the age of information, then privacy is the issue of our times

A. Acquisti, L. Brandimarte, and G. Loewenstein. Science, 2015

Ce chapitre possède des liens évidents avec les précédents. En effet, le respect de la vie privée englobe la protection des données personnelles, qui peut être reliée à leur sécurité, via la garantie de leur confidentialité et leur intégrité notamment. D'autre part, si les individus sont confiants dans le niveau (perçu) de protection de leurs données par un système (d'authentification biométrique), cette confiance contribue indiscutablement à l'utilisabilité de ce système.
Ce chapitre est organisé de la façon suivante. Une première partie permet d'appréhender la notion de vie privée selon différents points de vue (historique, éthique, informatique). Une deuxième partie est consacrée à la réglementation sur le respect de la vie privée : en tant que droit fondamental, puis à travers le *Privacy by Design* et le Règlement Général européen sur la Protection des Données (RGPD), avec un accent sur son application dans le cadre de la recherche. La dernière partie est centrée sur les techniques de biométrie révocable, plus précisément sur l'algorithme populaire du BioHashing, qui est au cœur de la thèse de Rima Belguechi [Bel15].

1 Introduction : différents points de vue sur la notion de vie privée

On peut distinguer deux types d'approches pour définir la protection de la vie privée : une première, qui tente de **protéger l'identité de l'utilisateur**, et une seconde qui se concentre sur la **protection des données de l'utilisateur**. Dans la première approche, la protection de la vie privée est obtenue en empêchant un adversaire de faire le lien entre les données correspondant à un individu et son identité. Dans le second cas, en revanche, l'identité du propriétaire des données est supposée être connue de l'adversaire. La protection des données est alors obtenue en modifiant les informations divulguées à l'adversaire, en les transformant, par exemple en les remplaçant par une valeur perturbée, en augmentant leur granularité, en y ajoutant des résultats fictifs, etc. On trouve des propositions pour les deux approches dans le cadre de la

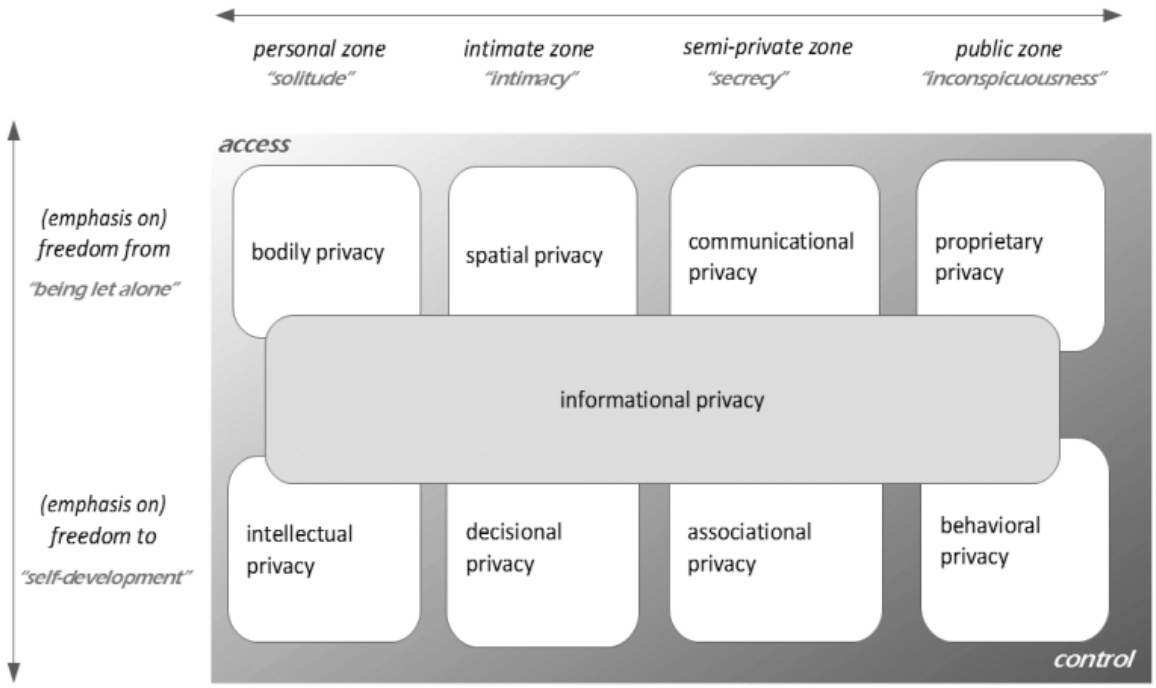


FIGURE 3.1 – Typologie de la vie privée informationnelle [Koo+16]

protection des données de localisation, particulièrement sensibles, dans la thèse de Nicolás E. Bordenabe [Bor14].
D. Solove propose une définition pratique de la notion de vie privée [Sol21] :
« *When people want privacy, they don't want to hide away their information from everyone ; instead, they want to share it selectively and make sure that it isn't used in harmful ways. Privacy isn't all-or-nothing – it's about modulating boundaries and controlling data flow.* »

La notion même de vie privée possède de nombreuses acceptions, qui toutes dépendent d'un contexte historique, social, culturel, légal. On distingue plusieurs types de vie privée. Chaque type est centré sur un aspect particulier : sur le corps (*bodily privacy*, correspondant à l'intégrité physique), sur les lieux de vie (*locational privacy*, correspondant à la protection du foyer, de la maison), sur les relations inter-personnelles (*relational privacy*, correspondant à la protection de la vie de famille, de la vie sociale) et enfin sur les données (*informational privacy*, correspondant à la protection des données personnelles ainsi qu'au secret des correspondances). Plus de détails sur ces différentes catégories peuvent être retrouvés dans le livre de B. Roessler [Roe05].
La part de la vie privée impliquée dans l'authentification biométrique est par nature en lien avec sa dimension informationnelle. Une typologie plus approfondie de la vie privée au sens de la protection des données personnelles est illustrée par la figure 3.1, extraite de la référence [Koo+16].
De façon générale, un défi se pose lorsqu'on réfléchit à la question : comment mesurer le respect de la vie privée dans un système ? L'article de Wagner et Eckoff [WE18] propose un tour d'horizon plus que complet des différentes métriques pour évaluer la protection de la vie privée. Ces métriques sont classées selon quatre caractéristiques :

1. *Les modèles de l'attaquant* décrivent les capacités supposées des adversaires ;
2. *Les sources de données* – données publiques, données observables, données ré-affectées – décrivent la manière dont l'attaquant peut obtenir des informations que le système

respectueux de la vie privée est censé protéger ;

3. *Les entrées* précisent quelle information est utilisée pour calculer une métrique : l'estimation de l'attaquant, les ressources dont il dispose, le véritable résultat, les connaissances préalables ou encore les paramètres ;
4. *Les sorties* décrivent les propriétés qui sont évaluées par les métriques de confidentialité. La taxonomie de la figure 3.2 présente huit catégories : l'incertitude ; le gain ou la perte d'information ; la similarité des données ; l'indiscernabilité ; la probabilité de succès de l'attaquant ; l'erreur ; le temps ; l'exactitude/la précision.

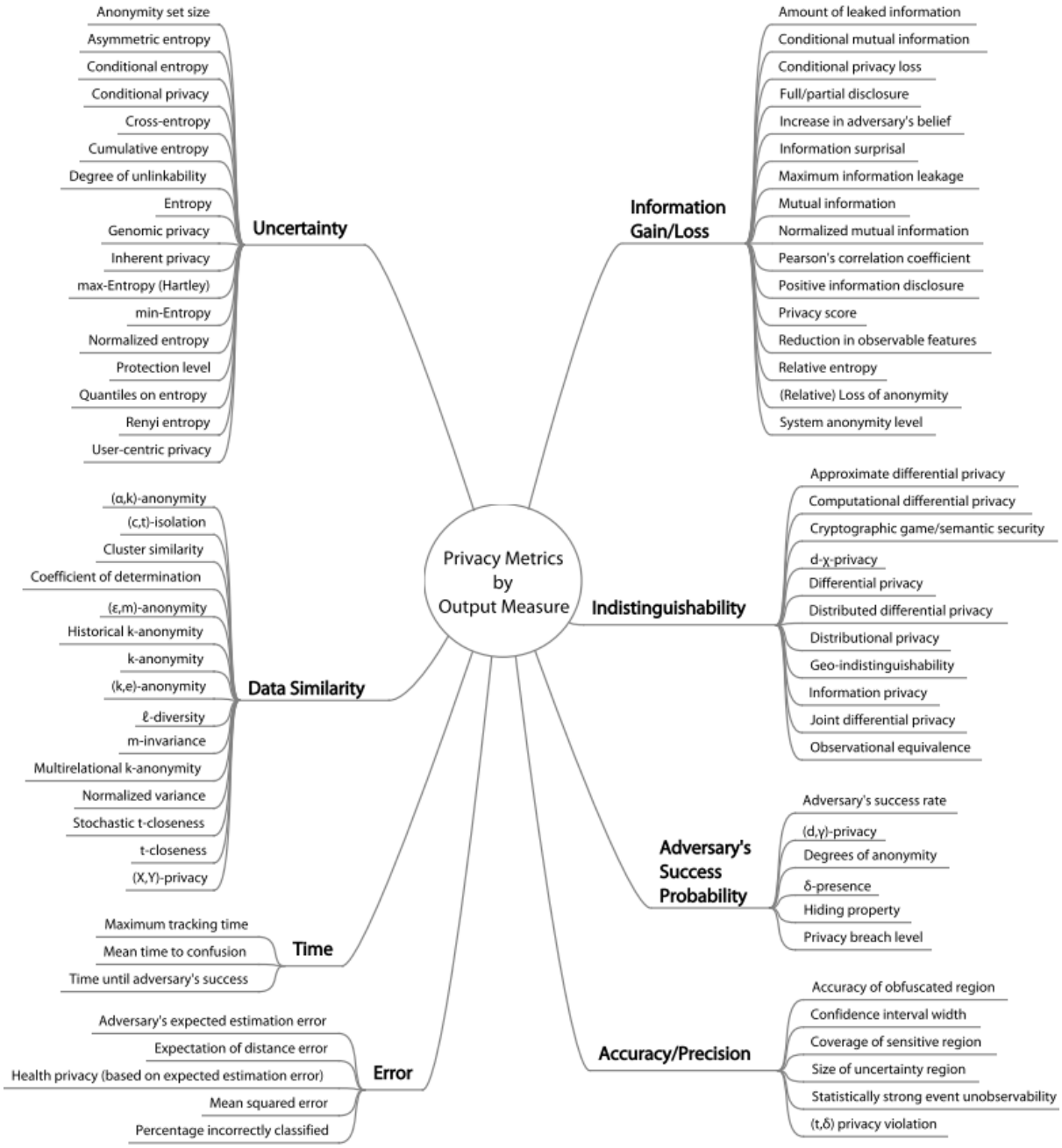


FIGURE 3.2 – Taxonomie des métriques de privacy, en fonction des propriétés évaluées [WE18]

La vie privée est souvent définie par opposition à la vie publique. L'étude de cette notion déborde largement du cadre législatif. En effet elle intéresse de nombreux domaines, comme la sociologie, la médecine ou l'histoire, en passant par l'économie, l'éthique, et l'informatique bien sûr. Cette diversité de traitements rend quasi impossible l'établissement d'une définition universelle. On peut noter toutefois deux termes qui semblent constituer un socle commun, à savoir l'*accès* et le *contrôle*, comme on le voit sur la figure 3.1.

Dans le cadre de ce mémoire, il m'est impossible de traiter la pluralité des définitions de la vie privée, aussi ai-je choisi de présenter trois éclairages sur cette notion complexe et polysémique : tout d'abord à travers l'évolution historique du concept de vie privée, puis en présentant des considérations éthiques sur ce sujet, pour terminer par un focus sur la vision informatique.

1.1 Point de vue historique

Une mauvaise compréhension de l'histoire de la notion de vie privée peut mener à des peurs non fondées, uniquement dues à une interprétation erronée. Cette partie retrace en quelques dates et faits marquants les variations de la notion de vie privée au cours des siècles.

Avant 1500, la notion même d'individualité n'existe pas. Seule la distinction entre la sphère familiale (donc privée, *oikos* en grec) et la sphère publique (principalement les activités politiques, *polis* en grec), développée par Aristote (384-322 av. JC) au IV$^{\text{ème}}$siècle av. JC, existe.
A partir du XVI$^{\text{ème}}$siècle, jusqu'à la Révolution Française, le respect de la vie privée recouvre l'acception suivante : autonomie personnelle et individualité. L'émergence d'une classe moyenne induit le développement du travail intellectuel ainsi que le choix de son espace de vie. Ce nouveau sens de l'autonomie, cette nouvelle conscience de soi-même sont permis grâce à l'invention de l'imprimerie, qui rend possible l'essor de la lecture personnelle : lire nécessite de disposer d'un endroit calme, privé, retiré du reste de la cellule familiale, pour laisser vagabonder son esprit, un lieu propice à la réflexion sur le livre entamé, mais aussi sur le monde, la condition humaine, etc. Même si la notion de vie privée revêt une connotation négative chez Shakespeare, en raison de son lien avec la solitude, ou encore avec les conspirations politiques, le philosophe anglais John Locke (1632-1704) – un des pères de l'*État de droit* – met à jour le but légitime des individus à protéger leur vie privée, leur liberté et leur propriété.
Puis au XIX$^{\text{ème}}$siècle, période d'urbanisation intense, de développement de la communication et de formation des états, des évolutions importantes de la société ont lieu, comme l'augmentation notable de la population, qui fait émerger un besoin de contrôle par le pouvoir politique. Les nouvelles technologies de l'époque sont en plein essor : le courrier postal, le télégraphe, le téléphone, les enregistrements audio, la photographie. L'augmentation du niveau d'éducation favorise quant à lui la lecture, l'écriture, l'échange de lettres, qui contribuent à la naissance du journalisme moderne et... de la presse à scandales. Certains personnages publics développent alors un intérêt tout particulier à la protection de leur vie privée. Le libéralisme fait surgir la liberté d'opinion, ainsi que de nouvelles lois de protection des droits individuels. L'article juridique *The right to privacy* [WB90], publié en 1890 par Warren et Brandeis, juristes américains, constitue toujours une référence en matière de droit à une vie privée pour chacun, comme le montre cette citation :
« *Le fait que chaque individu doit avoir une protection complète tant de sa personne que de ses biens est un principe aussi vieux que la loi commune ; mais il a semblé nécessaire de le redéfinir périodiquement pour renouveler sa nature exacte et l'étendue d'une telle protection.* »
Le XIX$^{\text{ème}}$siècle voit donc émerger deux visions de la notion de vie privée : (i) celle de l'état de surveillance, omniscient ; (ii) celle du citoyen, en tant qu'idéal et aspiration de chacun. La

première menace directement la vie privée des citoyens, dans le sens où la discipline règne, avec comme point d'orgue le panoptique de Bentham (1748-1832) : il s'agit d'une architecture carcérale, plutôt de forme circulaire, où toutes les cellules individuelles peuvent être surveillées par un seul gardien placé dans une tour au centre. La seconde vision, celle des citoyens, fait suite aux différentes Révolutions (française, américaine) et Déclarations des Droits (Bill of Rights en 1689 en Angleterre et Déclaration des Droits de l'Homme et du Citoyen en France en 1789).
Cette tendance se confirme au XX$^{\text{ème}}$siècle, pendant lequel le respect de la vie privée constitue une attente plus fondamentale et intellectuelle. En réaction aux deux Guerres Mondiales et à la fin du colonialisme, les libertés et les droits fondamentaux sont réaffirmés. La Déclaration Universelle des Droits de l'Homme est ainsi proclamée par l'Assemblée générale des Nations Unies le 10 décembre 1948. Elle revêt uniquement une valeur déclarative. L'article 12 concerne la vie privée :
« *Nul ne sera l'objet d'immixtions arbitraires dans sa vie privée, sa famille, son domicile ou sa correspondance, ni d'atteintes à son honneur et à sa réputation. Toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes.* »
Un traité international est ensuite signé le 4 novembre 1950 par les États membres du Conseil de l'Europe et entre en vigueur le 3 septembre 1953 : la Convention Européenne des Droits de l'Homme (CEDH, aussi appelée Convention de sauvegarde des droits de l'Homme et des libertés fondamentales). Afin d'assurer le respect des engagements de la CEDH, une juridiction internationale est instituée en 1959 à Strasbourg : la Cour Européenne des Droits de l'Homme.
Le respect de la vie privée de chacun (hommes, femmes, enfants) se voit également renforcé par l'émergence de l'État-Providence : par exemple, la protection sociale impose des normes pour les logements, avec des pièces supplémentaires pour respecter la tranquillité de chacun. En fait, l'apparition du domaine social constitue un phénomène totalement nouveau, car il ne relève ni du domaine public, ni du domaine privé. Il s'agit d'un espace commun, celui d'un partage de biens, de valeurs ou de convictions. D'où un bouleversement dans l'opposition entre ces deux sphères, opposition constatée et admise depuis l'Antiquité (*cf.* Aristote, au début de ce paragraphe). Le roman 1984 de G. Orwell [1] pousse plus loin la vision d'une société où les libertés individuelles seraient totalement réduites : le terme « *Big Brother* », outil implacable d'un régime policier et totalitaire, est passé dans le langage courant ; il est aujourd'hui encore synonyme d'une société de la surveillance.
A partir des années 1970, la vie privée s'entend à l'ère des ordinateurs comme la protection, la confidentialité des informations (*information privacy*). Les différents cadres réglementaires européens ou français seront détaillés dans la partie suivante (voir page 82). Depuis les attaques du 11 septembre 2001 et la lutte contre le terrorisme qui s'est renforcée ensuite, la vie privée est un droit fondamental que les États s'autorisent à rogner sous prétexte de renforcer la sécurité des citoyens. Contre l'argument « personne ne peut être contre les mesures de sécurité s'il n'a rien à cacher, rien à se reprocher », Edward Snowden, lanceur d'alerte américain rétorque [2] :
« *Dire que votre droit à la vie privée importe peu car vous n'avez rien à cacher revient à dire que votre liberté d'expression importe peu, car vous n'avez rien à dire. Car même si vous n'utilisez pas vos droits aujourd'hui, d'autres en ont besoin. Cela revient à dire : les autres ne m'intéressent pas* ».
Pour aller dans le même sens, on peut retrouver une démonstration dans l'article de Solove [Sol07] que l'argument de l'autorité étatique précédent n'est pas recevable, car il s'appuie sur une définition très réduite – donc réductrice – de la vie privée. Il y a également une asymétrie entre le droit à la vie privée des citoyens (sans besoin de justification), et le devoir de transparence des États vis-à-vis de la collecte de données par exemple, comme l'exprime l'expression de Julian Assange (2007) : « *privacy for the weak and transparency for the powerful* ».

1. 1984 (George Orwell, 1949)
2. *Nothing to hide*, Documentaire produit par Mihaela Gladovic et Marc Meillassoux, 2017

Il apparaît donc que la notion même de vie privée est un concept fluctuant au cours de l'histoire, en constante évolution donc, qui ne peut être étudié hors d'un contexte social, politique, économique et international.

1.2 Point de vue éthique

Comme mentionné dans le livre édité par Bart van der Sloot et Aviva de Groot [GS18, chapitre 5], un point de vue éthique sur la vie privée permet de répondre aux questions suivantes : quelle est la valeur de la vie privée ? quelles normes concernant la vie privée doivent être respectée par les individus, la société, l'état ?
Même si certains chercheurs renoncent à figer une définition précise – forcément réductrice – de la notion de vie privée, un grand nombre de définitions incluent les termes *accès* et *contrôle*. Tout d'abord on peut citer la définition de Reiman [Rei95] : « *privacy is the condition in which others are deprived of access to you* ». Le terme *accès* signifie soit un accès physique à une personne, soit un accès à des informations sur cette personne. Le point de vue éthique s'intéresse alors (i) à la manière dont cet accès est obtenu ; (ii) à quoi exactement on a accès. Cependant Fried [Fri84] a montré que l'*accès*, seul, ne suffit pas. Il faut lui adjoindre le *contrôle*, ceci pour éviter le cas absurde d'un homme isolé sur une île déserte : effectivement, personne n'a accès à lui, mais la notion de vie privée n'a aucun sens dans ce contexte. La définition de Fried relie les deux notions : le respect de la vie privée consiste à contrôler l'accès à soi. En fait, l'accès à une personne ou à des informations personnelles n'est pas à rejeter en bloc : l'important est de pouvoir contrôler cet accès.

On peut remarquer que, d'une part, ces définitions se réfèrent à l'article *The Right to Privacy* de Warren et Brandeis [WB90] ; d'autre part, elles mettent en exergue la question éthique suivante : est-ce que les protections légales qui existent sont suffisantes pour protéger les droits individuels, en particulier le contrôle exercé par chaque individu sur sa vie privée ? Cette question est d'autant plus cruciale que les technologies actuelles tendent à défier les normes sociales. Les travaux de Beate Roessler [GS18, p. 137] proposent une réponse selon trois dimensions de la vie privée :

1. La dimension locale : comment chacun contrôle l'accès à ses espaces physiques (par exemple, la clé sur la porte d'entrée de sa maison) ;
2. La dimension informationnelle (le terme *informations* renvoie souvent aux *données*) : contrôle sur ce que les autres peuvent savoir de nous ;
3. La dimension décisionnelle : contrôle sur l'accès (symbolique) à nos décisions personnelles, nos valeurs, nos objectifs, les raisons de nos décisions, etc.

Selon B. Roessler [Roe05], ces trois dimensions de la vie privée doivent actuellement faire face à des défis.

Défis de la dimension locale

Les objets connectés sont par définition connectés à Internet, sphère publique et totalement ouverte. Simultanément, ils envahissent notre espace privé, ce qui n'est pas inquiétant en soi. Ce qui est potentiellement dangereux, c'est la collecte, le stockage, l'analyse, etc., à l'intérieur même de nos maisons, de nos données privées par des assistants personnels, des compteurs électriques ou autres radiateurs connectés, alors que notre espace physique privé était à peu près impénétrable jusque-là. Ces objets dits *intelligents* représentent également une menace potentielle pour les autres dimensions informationnelle et décisionnelle de notre vie privée.

Défis de la dimension informationnelle

Comme on vient de le voir, la collecte de données dans des domaines d'activité variés est de plus en plus importante. Couplée à des algorithmes d'apprentissage et de fouille de données toujours plus efficaces, cette collecte mène à la création de profils qui peuvent être utilisés dans différents contextes, ainsi nos identités (numériques) ne peuvent plus être cloisonnées. La persistance des traces informatiques constitue une autre source d'inquiétude : dès qu'une donnée est créée, aussitôt elle peut être partagée très facilement et exploitée dans un contexte inconnu, dans un but inconnu. Dans son best-seller *L'Âge du capitalisme de surveillance* [Zub19], Shoshana Zuboff retrace l'histoire de l'*économie de surveillance* : celle-ci repose sur un principe de subordination et de hiérarchie. Elle met en évidence le projet consistant à extraire une plus-value de nos agissements sur Internet. L'auteur dénonce que Google a découvert que nous avions moins de valeur que les pronostics que d'autres font de nos agissements :
« *L'industrie numérique prospère grâce à un principe presque enfantin : extraire les données personnelles et vendre aux annonceurs des prédictions sur le comportement des utilisateurs. Mais, pour que les profits croissent, le pronostic doit se changer en certitude. Pour cela, il ne suffit plus de prévoir : il s'agit désormais de modifier à grande échelle les conduites humaines.* »

Défis de la dimension décisionnelle

Pour continuer sur la même thématique, le constat suivant est un doux euphémisme : le commerce de nos données, à travers l'essor du *Big Data*, de l'*IoT*, des réseaux sociaux augmente le risque que ces technologies nous influencent, manipulent nos comportements. En 2014, Facebook a lancé une étude sur la manipulation des émotions, sans le consentement de ses utilisateurs. D'un point de vue éthique, le problème n'est pas le sujet de l'étude, mais les conditions dans lesquelles elle a été réalisée, sans respecter les normes scientifiques et encore moins éthiques, comme le montre l'analyse de Flick dans son article [Fli16]. De la même façon, les média et réseaux sociaux aujourd'hui sont accusés de scléroser la pensée des internautes, en leur suggérant uniquement des contenus qui les intéressent déjà. On se retrouve piégé dans de véritables *bulles de filtres*. La curiosité intellectuelle n'est pas encouragée, impactant de fait la dimension décisionnelle de notre vie privée.

1.3 Point de vue informatique (TIC)

L'informatique fait partie des Technologie de l'Information et de la Communication (TIC) en permettant notamment le traitement, le stockage et la communication de données, y compris de données personnelles. La multiplication des collectes, des flux, des traitements de ces données personnelles impacte nécessairement la vie privée des individus. En fait, c'est la mise de relation entre des données capturées et l'identité des utilisateurs qui représente un risque pour la vie privée. La vie privée numérique est intrinsèquement liée à la sécurité numérique, dans le sens où elle relève de la confidentialité des données. Avec la prolifération des objets connectés, des applications mobiles, les risques sur la sécurité, et en particulier sur la vie privée, se multiplient. On est donc ramené au sujet plus général de la gestion des risques en sécurité informatique, sujet évoqué dans le paragraphe 2.2 du chapitre 1. Cependant, aucune méthode générique d'analyse des risques sur la vie privée n'existe, contrairement aux méthodes d'analyse des risques de sécurité. Il est donc très difficile d'évaluer le niveau de respect de la vie privée propre à un système, comme il est très difficile de comparer deux systèmes en matière de respect de la vie privée. L'article de Wagner et Eckoff [WE18] le prouve en recensant plus de quatre-vingt métriques disponibles sur le sujet.

Les données des utilisateurs peuvent être dans trois états, selon [GS18, chapitre 5] : (i) les données stockées, ou statiques (*i.e. data at rest*, par exemple dans une archive, ou une base de données peu consultée) ; (ii) les données utilisées, ou en cours d'utilisation (*i.e. data in use*, par exemple dans une base de données fréquemment modifiées par de multiples utilisateurs, ou des données en cours de traitement) ; (iii) les données en mouvement (*i.e. data in motion*, lorsque les données sont transférées via des réseaux informatiques à l'extérieur du système, ou sont utilisées par des services tiers). Les données ne sont pas soumises aux mêmes risques dans chacun des trois états. La suite de ce chapitre sera largement consacrée à la protection des données biométriques, données qui peuvent être dans l'un des trois états décrits précédemment.

Avec l'essor des ordinateurs personnels, puis le développement d'Internet et aujourd'hui l'essor des smartphones et des objets connectés dits *intelligents*, de véritables défis sont posés dans le domaine informatique, par la société ou la technologie, comme l'éthique de l'intelligence artificielle (dans les véhicules autonomes par exemple), l'éthique du « Big Data »(comment éviter les dérives dans les traitements de données massives), le développement de technologies respectueuses de la vie privées (ou *Privacy Enhancing Technologies*, PET).

On peut donc constater que le problème du respect de la vie privée est de plus en plus prégnant dans notre société : les nouvelles technologies représentent un défi toujours renouvelé pour les normes sociales. La question suivante se pose : est-ce que les protections légales existantes sont suffisantes pour protéger les droits des individus ? Le paragraphe suivant va donner quelques éléments de réponse.

2 La réglementation sur le respect de la vie privée

2.1 Le respect de la vie privée, un droit fondamental

La définition de la vie privée constitue sans doute le plus ancien principe juridique, en tant que séparation entre la sphère publique et la sphère privée. Son étymologie vient du latin *privare* (priver de), et par extension signifie soustraire quelque chose du domaine public. Pendant longtemps, la sphère publique a été gérée par le roi, ou le législateur de façon générale. La sphère privée, quant à elle, était gérée par le *pater familias*, le père de famille, qui régnait sur le foyer, le ménage. Le droit au respect de la vie privée était inclus dans les lois constitutionnelles de certains pays dès le XIII$^{\text{ème}}$siècle (en ce qui concerne l'intégrité physique, les lieux privés, les relations privées et le secret des correspondances). Au XX$^{\text{ème}}$siècle, le droit à la vie privée figure dans l'article 12 de la Déclaration Universelle des Droits de l'Homme des Nations Unies de 1948. Ce droit figure également dans les articles 7 (« *respect for private and family life* ») et 8 (« *protection of personal data* ») de la Convention Européenne des Droits de l'Homme de 1950 [3]. La Charte des droits fondamentaux de l'Union européenne, proclamée en 2000, comporte cinquante-quatre articles consacrant les droits fondamentaux des personnes, y compris les droits sociaux. Elle englobe le droit à la vie privée. Elle est contraignante pour les pays de l'UE dans le sens où elle se voit reconnaître une valeur constitutionnelle par le Traité de Lisbonne en 2007. En France, c'est le Code Civil qui reconnaît le droit à la vie privée : selon l'article 9 alinéa 1, « *Toute personne a droit au respect de sa vie privée* ». Comme le droit à l'image ou le droit à l'honneur, le droit à la vie privée fait partie des *Droits de la personnalité* : ce sont des droits inhérents à la personne humaine, qui la protègent en interdisant toute atteinte à ses droits les plus fondamentaux (sa vie, sa dignité, son corps).

3. https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/005

Dans la suite de cette partie, seuls les aspects informatiques de la vie privée ainsi que la protection des données personnelles seront considérés, qui concernent directement les systèmes biométriques. C'est au niveau européen que la réglementation évolue : tout d'abord, le principe de *Privacy by Design* (PbD) (ou respect de la vie privée dès la conception), puis le RGPD (Règlement Général sur la Protection des Données).

2.2 Privacy by design

Le *Privacy by Design* a été élaboré par Ann Cavoukian (Commissaire à l'information et à la protection de la vie privée de l'État d'Ontario) vers la fin des années 1990 au Canada. Ce concept, détaillé dans le document [Cav11], est une réponse à l'automatisation du traitement des données personnelles et à la quantité de plus en plus importante de données manipulées par les entreprises (ne nécessitant souvent pas une intervention humaine). Le *Privacy by Design* est une mesure préventive ayant donc pour but de limiter les risques d'abus et de violation des données : il implique une prise en compte de la protection de la vie privée des utilisateurs avant même la conception d'un système impliquant le traitement de données personnelles.

FIGURE 3.3 – Principes du *Privacy by Design* , selon [Cav11]

Le *Privacy by Design* énonce plusieurs principes, sept au total[4], comme illustré à la figure 3.3. Ces principes sont inspirés des :

1. Conception de mesures préventives et proactives
2. Protection par défaut (*Privacy by default*)
3. Prise en compte des règles sur la protection de la vie privée dans la conception des produits et durant leur utilisation ;
4. Protection optimale et intégrale ;
5. Sécurité de bout en bout – Protection du cycle de vie ;
6. Visibilité et transparence ;

4. https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf

7. Respect de la vie privée des usagers et/ou des cibles du service.

Le *Privacy by Default* est un principe qui vise tout produit ou service rendu public. Les standards en matière de protection des données personnelles s'appliquent alors par défaut, et ce sans recours à des procédures extérieures permettant cette protection. Le *Privacy by Default* est la garantie d'un niveau maximal de protection des données personnelles.
Les principes énoncés dans le *Privacy by Design* ont largement été repris dans le RGPD, présenté ci-dessous, ils ont évolué vers ce qu'on nomme *Data protection by design and by default*. La législation européenne sur la protection des données exige que les responsables de traitement et les sous-traitants intègrent les préoccupations en matière de protection des données dans chaque aspect de leurs activités de traitement. C'est cette approche qui est appelée *protection des données dès la conception et par défaut*. Elle est fondée sur le risque (c'est-à-dire qu'elle vise à minimiser les risques pour les personnes concernées) et exige une responsabilisation (c'est-à-dire que les organisations doivent être en mesure de démontrer comment elles se conforment à la législation).

2.3 Règlement Général sur la Protection des Données (RGPD)

En mai 2018, le Règlement Général sur la Protection des Données [Eur16] vient remplacer la Directive 95/46/CE du Parlement européen et du Conseil du 24 octobre 1995, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données. Selon le site web de l'UE [5], une Directive européenne est *un acte législatif qui fixe des objectifs à tous les pays de l'UE. Toutefois, chaque pays est libre d'élaborer ses propres mesures pour les atteindre.* La Directive de 1995 a donc été interprétée différemment dans chaque pays de l'UE, compromettant ainsi la construction d'un véritable cadre commun en matière de protection de la vie privée des citoyens européens, empêchant également l'harmonisation des politiques dans ce domaine, prérequis indispensable à une collaboration efficace à l'échelle européenne. Dans l'UE, un Règlement est défini comme *un acte législatif contraignant, qui doit être mis en œuvre dans son intégralité, dans toute l'Union européenne* [6]. C'est pourquoi il est si important que le RGPD soit, justement, un Règlement.

2.3.1 La protection des données à caractère personnel (ou données personnelles)

On a vu au début de ce chapitre que le contenu de la vie privée varie en fonction des périodes de l'histoire, il dépend également des pays, mais également de la personne (les contours de la vie privée d'une personnalité publique ne seront pas les mêmes que pour un citoyen ordinaire par exemple). En revanche, la définition d'une donnée personnelle est clairement définie par la CNIL [7] :

Définition 5 (Donnée à caractère personnel)**.** *Une donnée à caractère personnel est toute information se rapportant à une personne physique identifiée ou identifiable. Mais, parce qu'elle concerne des personnes, celles-ci doivent en conserver la maîtrise.*

On utilise indifféremment l'expression donnée personnelle ou donnée à caractère personnel.
Aujourd'hui, toujours selon la CNIL, une personne physique peut être identifiée :

- directement (exemple : nom et prénom) ;

5. https://europa.eu/european-union/eu-law/legal-acts_fr
6. *Ibid.*
7. https://www.cnil.fr/fr/definition/donnee-personnelle

- indirectement (exemple : par un numéro de téléphone ou de plaque d'immatriculation, un identifiant tel que le numéro de sécurité sociale, une adresse postale ou courriel, mais aussi la voix ou l'image).

L'identification d'une personne physique peut être réalisée :

- à partir d'une seule donnée (exemple : nom) ;
- à partir du croisement d'un ensemble de données (exemple : une femme vivant à telle adresse, née tel jour et membre dans telle association).

Historiquement, le premier instrument international juridiquement contraignant concernant la protection de données personnelles est la Convention du Conseil de l'Europe pour la protection des personnes à l'égard du traitement des données à caractère personnel (*Convention 108* du 28 janvier 1981). Elle a été modernisée en 2018 (*Convention 108+*), en lien avec le RGPD.

Il ne s'agit pas ici d'analyser le texte complet du RGPD, mais plutôt de mettre en lumière quelques éléments clés. Le principe général de ce règlement est de renforcer la protection des données à caractère personnel, à travers plusieurs points [8] :

- Consentement clair requis pour traiter les données ;
- Limitation du recours au traitement automatisé pour arrêter des décisions, par exemple dans le cas du profilage ;
- Droit de rectification et de suppression des données collectées lorsque la personne concernée a le statut d'enfant, y compris le droit à l'oubli ;
- Droit à une notification en cas de violation des données ;
- Information plus complète et claire concernant le traitement ;
- Droit de transférer les données d'un prestataire à un autre ;
- Accès plus aisé aux données à caractère personnel ;
- Garanties plus rigoureuses en cas de transfert de données à caractère personnel hors de l'UE.

Parmi les données personnelles, certaines données sont considérées comme particulièrement sensibles. Le RGPD interdit de recueillir ou d'utiliser ces données, sauf, notamment, si la personne concernée a donné son consentement exprès (démarche active, explicite et de préférence écrite, qui doit être libre, spécifique, et informée). Ces exigences concernent les données suivantes :

- les données relatives à la santé des individus ;
- les données concernant la vie sexuelle ou l'orientation sexuelle ;
- les données qui révèlent une prétendue origine raciale ou ethnique ;
- les opinions politiques, les convictions religieuses, philosophiques ou l'appartenance syndicale ;
- les données génétiques et biométriques utilisées aux fins d'identifier une personne de manière unique.

En tant que donnée personnelle sensible, toute donnée biométrique devrait donc être protégée et ne devrait pas être stockée, ni transiter entre deux éléments d'un système biométrique sous sa forme brute, ni sous une forme qui permet de reconstruire (une partie de) la donnée originale. Cette problématique fera l'objet du paragraphe 3.

En tant que RIL (Relais Informatique et Libertés) pour la recherche au sein du laboratoire GREYC, j'ai pour mission de recenser et accompagner les demandes de traitement de données déposées par mes collègues, avant de passer le relais au DPO (Data Protection Officer) d'une des tutelles du laboratoire. Cette mission me permet d'envisager les thématiques liées à la vie privée et la protection des données sous un angle complémentaire à mes propres travaux. Je m'intéresse donc plus particulièrement à l'application du RGPD dans le cadre de travaux de

8. https://www.consilium.europa.eu/fr/infographics/data-protection-regulation-infographics/

recherche, cadre qui présente quelques différences fondamentales par rapport au cadre général.

2.3.2 RGPD et biométrie

La recherche en biométrie est particulièrement concernée par l'article 35 du RGPD, qui introduit la notion d'Analyse d'Impact relative à la Protection des Données (AIPD, ou DPIA - *Data Protection Impact Assessment*). Cette analyse doit être réalisée dès lors qu'un traitement sur des données est susceptible d'engendrer un risque élevé pour les droits et libertés des personnes concernées. Cette étude d'impact doit faire apparaître les caractéristiques du traitement, les risques et les mesures adoptées. Tous les détails sont disponibles dans la publication des Autorités de protection des données européennes (le G29) [17]. Dès l'instant où une collecte de données biométriques est programmée, il est donc impératif de lancer une AIPD. Toutefois, il est important de souligner que les données biométriques, dans un contexte de recherche publique, ont un statut spécial : le RGPD prévoit quelques options, offrant des latitudes aux pays membres de l'UE, notamment en ce qui concerne les données de recherche. La France a choisi d'assouplir les contraintes, pour coller aux mieux aux objectifs de recherche. Ainsi, la durée de conservation peut être allongée, la finalité peut évoluer au cours du temps... On trouve sur le site de la CNIL un document [9] intitulé *Présentation du régime juridique applicable aux traitements poursuivant une finalité de recherche scientifique (hors santé)* :
« *Le RGPD définit la recherche scientifique largement. Son Considérant 159 indique ainsi que le traitement de données à caractère personnel à des fins de recherche scientifique devrait être interprété au sens large et couvrir, par exemple, le développement et la démonstration de technologies, la recherche fondamentale, la recherche appliquée et la recherche financée par le secteur privé . Un cadre particulier y est prévu pour ces traitements afin de concilier les spécificités de la recherche avec l'impératif de protection des données à caractère personnel.* »
Ce document fournit une liste de mesures appropriées à mettre en place pour pouvoir collecter des données biométriques dans un cadre de recherche :

- conformément à l'article 5.1-c) du RGPD, le principe de pertinence et de minimisation des données traitées doit être respecté ;
- l'anonymisation des données pourrait être mise en œuvre (le G29 a publié des recommandations à ce sujet en 2014) ; elle est obligatoire lors de la diffusion des résultats, diffusion nécessaire à la présentation des travaux de recherche (*cf.* l'article 116 du décret n°2019-536 du 29 mai 2019) ;
- la pseudonymisation doit être mise en œuvre toutes les fois où cela s'avérerait pertinent (sachant que les informations permettant de faire le lien entre les données et l'identité des individus doivent être conservées séparément et soumises à des mesures techniques et organisationnelles) ;
- une logique d'accès sécurisé et contrôlé doit être développée, en fonction de la sensibilité des données et des finalités des réutilisations futures ;
- la réalisation d'une analyse d'impact sur la protection des données est quasi systématique.

Pour aller plus loin dans la réflexion, l'article de Sanchez *et al.* [San+19] expose une adaptation de la façon de mener des recherches en biométrie en accord avec le nouveau cadre réglementaire européen. Les auteurs proposent une liste de contraintes inhérentes à toute collecte de données biométriques. Parmi ces contraintes, on peut citer : collecter l'identité et des informa-

9. https://www.cnil.fr/sites/default/files/atoms/files/consultation_publique_-_presentation_du_regime_juridique_applicable_aux_traitements_a_des_fins_de_recherche.pdf

tions pour contacter les utilisateurs ; collecter des informations supplémentaires (des habitudes par exemple) ; conserver les données sur une longue durée ; conserver le lien entre les données et un identifiant. Pour chaque contrainte, des explications ou des exemples sont donnés, ainsi que des mesures à mettre en place pour garantir la conformité avec le RGPD. Les auteurs s'appuient sur une collecte d'empreintes digitales qu'ils ont réalisée auprès de six cents volontaires. On constate donc que l'adoption du RGPD a permis de clarifier et de structurer le cadre à respecter pour la collecte de données biométriques. La protection de ces données, en tant que données personnelles sensibles, demeure un enjeu scientifique.

2.4 Discussion

Même si le législateur a inévitablement un peu de retard sur les risques sur la vie privée dus aux nouveaux usages des smartphones et autres objets connectés, on constate que les protections des citoyens évoluent dans le bon sens, que ce soit au niveau du droit français, ou européen, comme le montre la progression du respect de la vie privée dans la hiérarchie des normes.

Dans son article [Kin17], Els Kindt pointe les insuffisances du RGPD en matière de protection des données biométriques, notamment à travers ce constat (Art. 4, définition 14 du RGPD [10]) : « *Les données biométriques sont les données à caractère personnel résultant d'un traitement technique spécifique, relatives aux caractéristiques physiques, physiologiques ou comportementales d'une personne physique, qui permettent ou confirment son identification unique*. »
Deux éléments sont donc nécessaires pour qu'une donnée biométrique entre dans ce cadre : le traitement spécifique et l'identification unique. Comme contre-exemple, E. Kindt propose une base de données de photographies d'empreintes ou de visages, simplement collectées et stockées : dans ce cas, ces images ne seraient pas des données biométriques (*cf.* le Considérant 51 du RGPD). Elle poursuit son analyse en étudiant la distinction entre quatre catégories de « données personnelles en lien avec des caractéristiques physiques, physiologiques ou comportementales d'un individu, et qui permettent son identification ou authentification » :

- *Les données personnelles ordinaires* (considérées comme non biométriques) : les données simplement stockées, les données de biométrie douce (qui ne permettent pas une identification unique). C'est le régime général du RGPD qui s'applique à ces données (Art. 6).
- *Les données biométriques* : cette catégorie, qui concerne les données biométriques en général (qui subissent un traitement spécifique et qui permettent ou confirment une identification unique) sont sujettes au même régime que les données personnelles ordinaires.
- *Les données biométriques sensibles* : ce sont les données qui sont traitées dans le seul but d'identifier de façon unique des individus. Dans ce cadre, l'article 9 du RGPD s'applique également : en principe, l'identification biométrique est interdite, sauf exceptions. Il faut souligner qu'il s'agit ici uniquement d'identification. L'authentification pourrait, sous couvert d'une certaine interprétation du RGPD, tomber dans la catégorie précédente (le but n'est pas une identification unique). Cependant, E. Kindt recommande des conditions supplémentaires : l'authentification devrait être transparente pour l'utilisateur, avec des garanties légales et techniques (pas de stockage dans une base centralisée), et le recours obligatoire à une AIPD.
- *Les données biométriques dont le but est d'identifier des individus à grande échelle* : ces données sont sujettes à des contraintes réglementaires additionnelles (Art. 35 et 36 du RGPD).

10. https://gdpr-text.com/fr/

L'article de Gellert [Gel18] porte sur la notion de risque, telle qu'elle est définie dans le RGPD. Il propose également une réflexion très pertinente sur la distinction entre PIA (Privacy Impact Assessment) et DPIA (Data Protection Impact Assessment). L'auteur rappelle que la vie privée va bien au-delà des données personnelles. Par conséquent mener un DPIA serait forcément réducteur par rapport à un PIA. Il s'appuie sur les travaux de R. Clarke [Cla11] :
A Data Privacy Impact Assessment is a study of the impacts of a project on only the privacy of personal data, whereas a PIA considers all dimensions of privacy.
Les dimensions de la vie privée auxquelles il fait référence sont : les informations personnelles, l'intimité de la personne, le comportement individuel, les communications personnelles, les pensées et les émotions. L'auteur, R. Gellert, dépasse ce clivage traditionnel et analyse en profondeur la notion de risque présente dans le RGPD. Il admet que le texte du règlement contient la définition de l'événement (à l'origine du risque), les conséquences (du risque) et un certain nombre de critères liés aux facteurs de risque. Il note toutefois qu'il manque un certain nombre d'éléments constitutifs d'un risque : (i) aucune mention n'est faite concernant la vraisemblance (ou probabilité d'occurrence) du risque (seul son niveau de gravité est mentionné), (ii) il y a peu de critères concernant les conséquences des risques : cela signifie que le règlement laisse une totale liberté quant au calcul des « risques relatifs aux droits et libertés des personnes concernées » dans la pratique, et (iii) le RGPD ne préconise aucune méthodologie pour réaliser un DPIA. La dernière phrase de l'article est la suivante :
« *Thus, to some extent, the way risk is defined and understood in the GDPR is pretty much irrelevant.* »

Els Kindt et Raphaël Gellert, auteurs des deux réflexions que j'ai choisi de présenter ici, sont des juristes. Leur point de vue externe au domaine de l'informatique permet une prise de recul nécessaire et salutaire dans mon travail de recherche, en élargissant les perspectives.
Finalement, on peut considérer que les données biométriques – qu'elles soient brutes, transformées, protégées – appartiennent à la catégorie des données personnelles sensibles. Le fait même qu'à partir d'une bribe de donnée biométrique, on pourrait être en mesure d'extraire une information nuisible pour la vie privée de l'utilisateur rend la protection de ces données indispensable. D'où le développement des techniques de biométrie révocable, qui vont être présentées dans la suite de ce chapitre, avec, comme toile de fond, les contributions de la thèse de Rima Belguechi [Bel15].

3 Biométrie révocable : définitions et contributions

La biométrie révocable fait partie d'un ensemble de techniques appelé *Biometric Template Protection* (BTP), ou schémas de protection biométrique. La taxonomie des différentes techniques est représentée à la figure 3.4, extraite du chapitre [SP17], qui propose une *revue de littérature systématique* et recense la plupart des travaux publiés sur les BTP (en 2017), pour faire émerger les verrous scientifiques sur le sujet. Après un rapide tour d'horizon des méthodes de biométrie révocable (*cancelable biometrics*), l'accent sera mis sur l'algorithme de BioHashing, entouré en rouge.

La norme ISO/IEC 24745 *Technologies de l'information – Techniques de sécurité – Protection des informations biométriques* publiée par le SC 27 en 2011 définit plus précisément *les exigences pour la protection des données personnelles pendant le stockage et le traitement des informations biométriques*. Ces contraintes se répartissent en deux catégories :

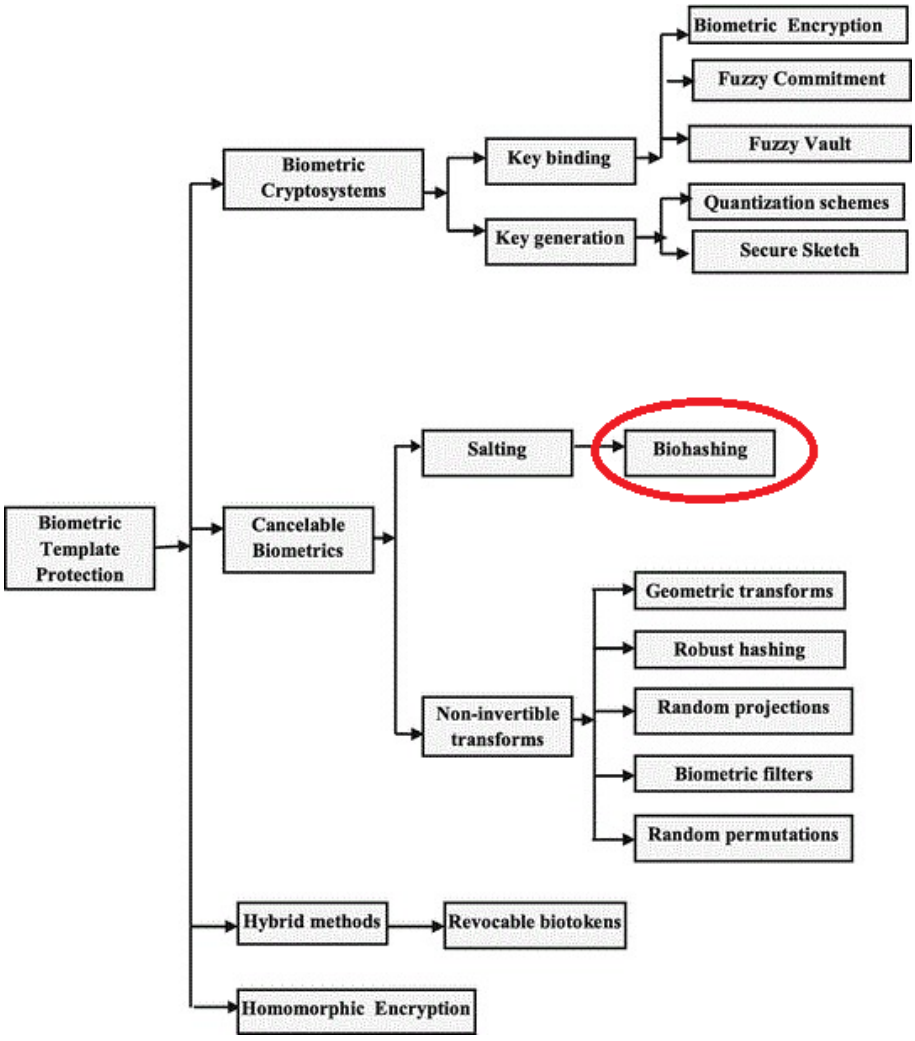


FIGURE 3.4 – Taxonomie des différents schémas de BTP, extraite de [SP17]

- **les exigences de sécurité** :
  - confidentialité : propriété qui protège les informations contre l'accès ou la divulgation non autorisés ;
  - intégrité : la propriété de sauvegarder l'exactitude et l'exhaustivité des données à protéger ;
  - renouvellement et révocabilité : la révocation est nécessaire pour empêcher l'attaquant d'accéder à l'avenir (ou dès à présent) à des données non autorisées.
- **les exigences de protection de la vie privée** :
  - non-inversibilité : pour empêcher l'utilisation des données biométriques à des fins autres que celles prévues à l'origine, les données biométriques sont transformées de façon irréversible avant d'être stockées ;
  - non-associativité : les références biométriques stockées ne doivent pas pouvoir être associées à d'autres applications ou bases de données ;
  - confidentialité : pour protéger les références biométriques d'un accès non autorisé, qui constitue un risque pour la vie privée, elles doivent rester confidentielles.

L'architecture globale d'un schéma de protection des données biométriques, recommandée par

la norme ISO/IEC 24745, est illustrée à la figure 3.5 [11].

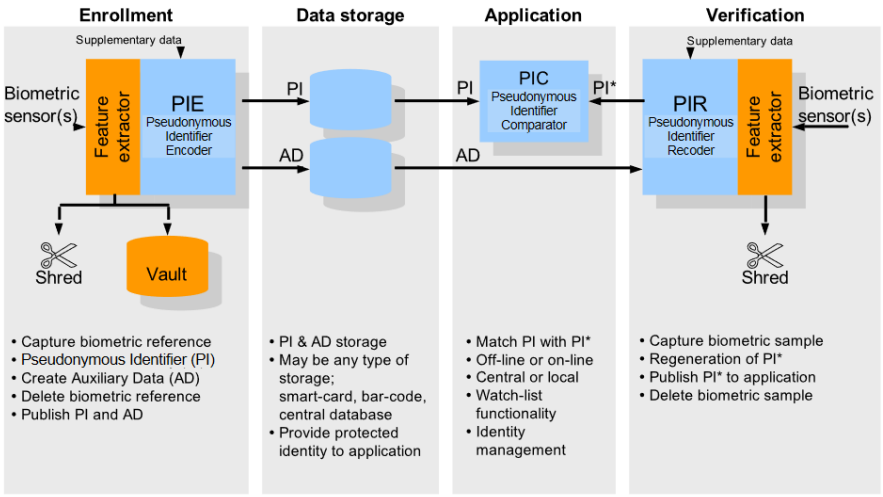


FIGURE 3.5 – Principe général d'un schéma de biométrie révocable, extrait de la norme ISO/IEC 24745

3.1 Biométrie révocable : définition, méthodes

La biométrie révocable constitue une des sous-familles de techniques de protection de la biométrie, comme on peut le voir à la figure 3.4.
La toute première mention à ce qui était appelé *biométrie indirecte* remonte à 1998, avec les travaux de Davida *et al.* [DFM98]. Trois ans plus tard, les articles de Ratha, Bolle et Connell [RCB01] et [BCR02] emploient l'expression *biométrie révocable*. C'est le début d'un engouement pour une thématique qui permet de contrer le principal argument en défaveur de la biométrie : la non-révocabilité intrinsèque de toute donnée biométrique. Le concept de *biométrie révocable*

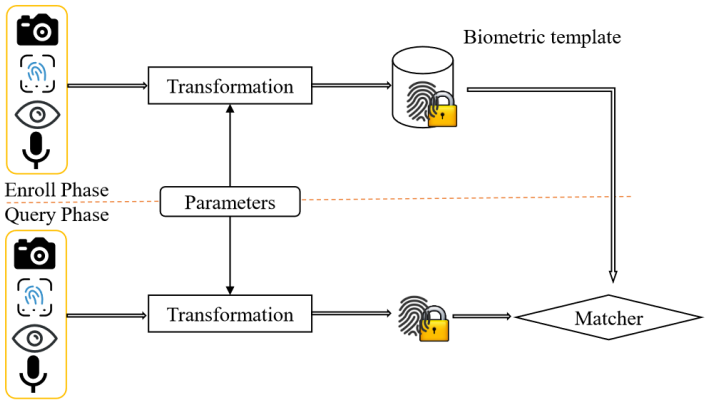


FIGURE 3.6 – Principe général d'un schéma de biométrie révocable

repose sur une transformation des données biométriques brutes (comme illustré à la figure 3.6), de telle sorte que les données transformées soient sûres et respectueuses de la vie privée, en

11. https://christoph-busch.de/files/Busch-EAB-ISO-24745-120713.pdf

accord avec les propriétés détaillées par Maltoni *et al.* [Mal+09] (certaines sont présentes dans la norme ISO/IEC 24745), également reprises dans l'article de Jain *et al.* [JNN13] :

- *Non-inversibilité :* il ne doit pas être possible de retrouver des informations sur la donnée biométrique originale.
- *Performance :* l'efficacité du système de vérification ne doit pas être détériorée par la transformation.
- *Diversité :* on doit pouvoir générer plusieurs données protégées à partir d'une seule donnée brute. Le recoupement de différentes données protégées ne doit pas affecter la protection de la vie privée.
- *Révocabilité :* on doit pouvoir facilement révoquer les données en cas de compromission.

L'article d'Inuma et Otsuka [IO13] détaille les expressions mathématiques, en terme de probabilités, de ces différentes propriétés, en utilisant le formalisme de la norme ISO/IEC 24745.

Les techniques de biométrie révocable permettent de ne jamais stocker les données originales : seules les données transformées sont conservées pour la vérification. La propriété de révocabilité est ainsi garantie : si une donnée transformée est compromise, il suffit de changer (les paramètres de) la fonction de transformation. La propriété de diversité est également assurée par le choix de fonctions différentes pour des applications distinctes. En outre, le système de vérification doit être sensible aux variations inter-classe (*i.e.* pouvoir distinguer deux utilisateurs différents) et en même temps robuste aux variations intra-classe (la donnée biométrique d'un utilisateur varie inévitablement, à cause de conditions de capture différentes, du vieillissement, etc.). Pour cela, les transformations de données biométriques utilisent une donnée ou clé secrète (les *paramètres* de la figure 3.6) en plus de la donnée biométrique originale. L'enrôlement consiste à calculer la transformée de la donnée de référence à l'aide de la clé, puis à stocker cette donnée transformée. La vérification nécessite le calcul de la transformée de la donnée présentée avec la clé de l'utilisateur, et la comparaison est effectuée entre les données transformées uniquement. On remarque également que des techniques standard (cryptographie, *differential privacy*, *random data perturbation techniques*, etc.) ne peuvent pas être appliquées directement aux données biométriques, en raison de leur variabilité.

On voit sur la figure 3.4 que la biométrie révocable se divise en deux branches : les techniques reposant sur un salage des données et les techniques reposant sur une transformation non inversible. Les schémas de salage appartiennent donc aux schémas de protection de la biométrie et reposent sur une transformation des données par une fonction inversible (le processus global restant non inversible). La clé (ou les paramètres) doit être stockée de façon sécurisée, ou être mémorisée par l'utilisateur (s'il s'agit d'un mot de passe, d'un code PIN, etc.). Nous avons démontré dans l'article [LCR13] qu'il est absolument nécessaire que la clé soit stockée indépendamment de la donnée transformée : il s'agit d'une authentification forte, à deux facteurs (la donnée biométrique et la clé), par conséquent il faut que les deux facteurs ne soient liés d'aucune façon. Le schéma de BioHashing est la technique de salage la plus populaire depuis une vingtaine d'années. Il est au cœur de la thèse de Rima Belguechi [Bel15], intitulée *Sécurité des systèmes biométriques : révocabilité et protection de la vie privée*, soutenue en 2015. Cette thèse a été co-dirigée par Christophe Rosenberger et par Samy Ait-Aoudia (Ecole NAtionale Supérieure d'Alger), je l'ai encadrée à partir de mon changement d'équipe en janvier 2011.

3.2 Algorithme de BioHashing

Dans le chapitre [SP17], déjà cité, Sandhya et Prasad présentent un état de l'art de la littérature publiée sur les techniques de salage pour la biométrie révocable depuis 2001. Les premiers travaux concernent la paume de la main [Con+05], l'empreinte digitale [TNG04], le visage [TGN06]. Le BioHashing a été décrit par Andrew B.J. Teoh, David C.L. Ngo et Alwyn Goh, dans plusieurs articles [GN03], [TNG04], [TKL08].

Principe du BioHashing

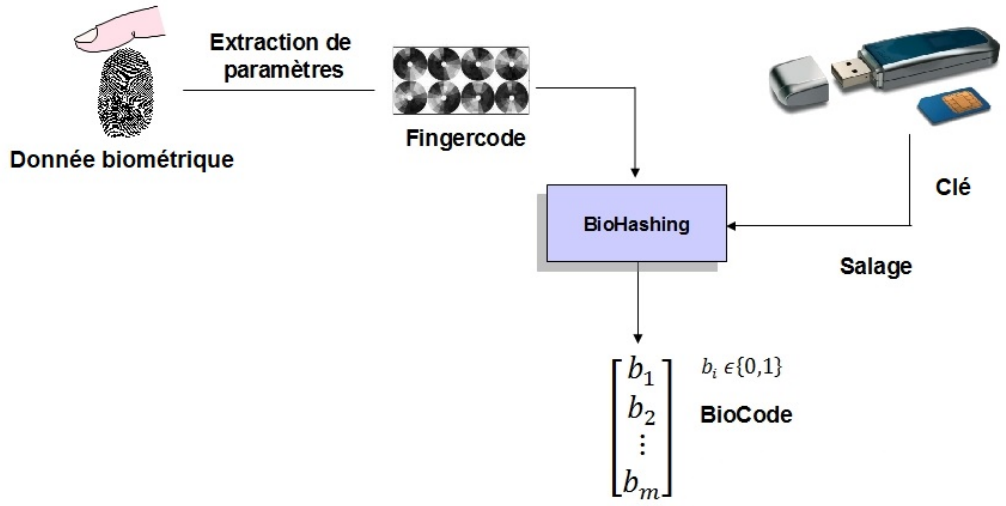


FIGURE 3.7 – Principe général de l'algorithme de BioHashing , inspiré de [TNG04]

Le BioHashing repose sur deux facteurs : une donnée biométrique et une clé (voir la figure 3.8). Dans la thèse de Rima Belguechi, la modalité étudiée est l'empreinte digitale, aussi vais-je conserver cette modalité comme illustration.
Lors de la phase d'enrôlement, l'utilisateur présente son empreinte et la clé (secrète) stockée sur une clé USB, une carte à puce, ou plus généralement un *token*. Des paramètres sont extraits de l'empreinte sous forme de FingerCode, gabarit non transformé. La fonction de transformation prend comme entrée ce FingerCode et la clé secrète pour générer un BioCode binaire.
L'algorithme de BioHashing lui-même comporte deux étapes :

- une projection du FingerCode sur une matrice aléatoire orthonormée ;
- une quantification pour obtenir le vecteur binaire appelé BioCode.

Plus précisément, on considère un gabarit d'empreinte, le FingerCode $F = [f_1 \dots f_n]$ de dimension n. D'un autre côté, à l'aide de la clé, on génère m vecteurs aléatoires linéairement indépendants de dimension n, que l'on peut regrouper sous la forme d'une matrice aléatoire $R = [r_{i,j}]_{i=1,n,j=1,m}$. On peut utiliser un générateur de loi uniforme, ou gaussienne normale centrée, ou encore d'autres lois de probabilité, comme étudié par Achlioptas dans son article [Ach03]. Cette matrice est ensuite orthonormalisée par l'algorithme de Gram-Schmidt.

Théorème 1 (Orthonormalisation de Gram-Schmidt)**.** *Soit $\{u_k\}_{k=1,n}$ une famille libre d'un espace vectoriel pré-Hilbertien E muni d'un produit scalaire $\langle , \rangle$.*
Alors il existe une unique famille $\{v_k\}_{k=1,n}$ telle que :

- *$\{v_k\}_{k=1,n}$ est une famille de vecteurs orthonormaux*

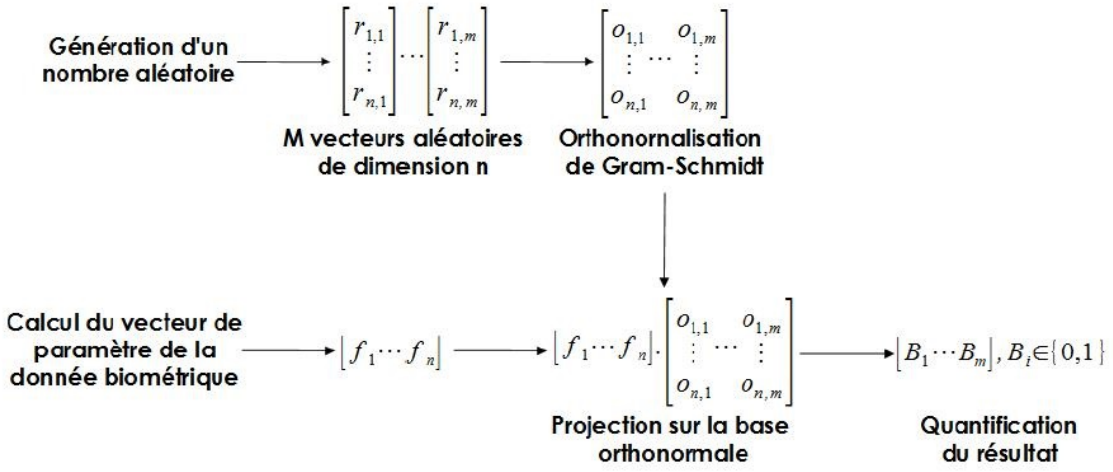


FIGURE 3.8 – Détails de l'algorithme de BioHashing , extrait de [Bel15]

- $\forall\ 1 \leq k \leq n$, $\mathrm{Vect}(v_1, \ldots, v_k) = \mathrm{Vect}(u_1, \ldots, u_k)$

La famille $\{v_k\}_{k=1,n}$ est appelée famille orthonormalisée de $\{u_k\}_{k=1,n}$ par le processus de Gram-Schmidt.

En pratique, on définit $\{v_k\}_{k=1,n}$ par récurrence :

- $v_1 = \dfrac{u_1}{\|u_1\|}$
- for $k \geq 2$, $v_k = \dfrac{m_k}{\|m_k\|}$ avec $m_k = u_k - \langle u_k, v_1\rangle v_1 - \langle u_k, v_2\rangle v_2 - \ldots - \langle u_k, v_{k-1}\rangle v_{k-1}$

La matrice résultante est notée O. L'étape suivante consiste à projeter le FingerCode sur l'espace de dimension n en le multipliant par la matrice $O = [O_1 \ldots O_m]$, les $O_j, j = 1, m$ étant les vecteurs colonnes de O. On note $W = FO$. Jusque-là, la première partie du BioHashing a permis de mélanger l'information contenue dans les données biométriques. Cependant, cette simple projection aléatoire par une matrice orthonormée est totalement inversible. La dernière étape de quantification est par conséquent indispensable : on calcule le BioCode $B = [b_1 \ldots b_m] = \mathrm{Sgn}\left(\sum FO_j - \tau\right)$, où τ est un seuil de quantification (il peut être choisi égal à zéro, à la moyenne, ou à la médiane des coefficients de W). Le BioCode est donc un vecteur binaire, de dimension m.
La vérification consiste à comparer le BioCode stocké comme référence avec le BioCode calculé lors de la requête, à partir de l'empreinte et de la clé présentées au système. La comparaison repose sur le calcul d'une distance de Hamming entre ces deux BioCodes.
Les résultats obtenus dans la thèse de Rima Belguechi sont les suivants, sur la base d'empreintes FVC2002-DB2, avec une extraction des caractéristiques de l'empreinte par un banc de filtres de Gabor :

- sans protection : EER = 10.25%,
- avec BioHashing, dans le cas idéal (sans vol de clé) : EER = 0%
- avec BioHashing, dans le cas où la clé est volée (*stolen token scenario*) : EER = 11.40%, avec un seuil de quantification égal à la médiane.

On constate que les performances avec vol de la clé sont au même niveau que celles du système biométrique non protégé, ce qui semble assez normal. Ce point est abordé dans l'article de Jin *et al.* [Jin+18]. Les premiers articles évoquant les scénarios de vol de la clé ou vol de la biométrie

n'étaient pas clairs sur ce sujet : ils considéraient des seuils de décision (pour la vérification) différents en fonction des attaques, alors qu'un système réel possède un seuil fixé, identique pour toutes les expériences. Dans sa thèse, Rima Belguechi a choisi de fixer ce seuil en tenant compte du scénario de vol de la clé, plus précisément pour limiter son impact.

L'algorithme de BioHashing fait partie des méthodes de salage préservant les distances, ou les similarités, en relation avec l'algorithme de Johnson-Lindenstrauss. Dans certains travaux [Che+19], [Jin+18], elles sont rattachées aux méthodes de hachage sensible à la localisation (Locality Sensitive Hashing, LSH), adaptées à la fouille de données en grande dimension, qui peuvent notamment prendre la forme de projections aléatoires.

Le lemme de Johnson-Lindenstrauss

C'est ce lemme, au cœur du BioHashing, qui garantit la propriété de conservation des distances. Cette propriété paraît essentielle dans la prise en compte des variabilités des données biométriques. Pourtant, on verra un peu plus loin qu'elle représente une réelle faiblesse des schémas à base de projection aléatoire (*cf.* [WP10], ou [Che+19] par exemple).

Le lemme de Johnson-Lindenstrauss (JL) établit qu'un ensemble de points dans un espace de grande dimension peut être plongé dans un espace de dimension bien inférieure tout en conservant approximativement les distances entre les points, deux à deux. La valeur de la dimension de l'espace d'arrivée dépend du nombre de points dans l'ensemble initial et du facteur d'approximation requis pour les distances entre les points. Le lemme est énoncé dans l'article de référence de Johnson et Lindenstrauss [JL84] :

Lemme 1. *Soit $\epsilon \in]0,1[$ un réel et Q un ensemble de $\#(Q)$ points dans $\mathbb{R}^N$. Soit n un entier positif tel que $n > \mathcal{O}\left(\frac{\ln(\#(Q))}{\epsilon^2}\right)$. Alors il existe une application lipschitzienne $f : \mathbb{R}^N \to \mathbb{R}^n$ telle que, pour tout $u, v \in Q$:*

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2 \tag{3.1}$$

Dans l'article de Gupta et Dasgupta [GD99], la borne sur le paramètre k est précisée : $n \geq \frac{4\ln(\#(Q))}{\epsilon^2/2 - \epsilon^3/3}$.

On comprend pourquoi ce lemme est adapté aux données biométriques. Il permet de résoudre les deux conditions qui semblaient irréconciliables, à savoir : être robuste à la variabilité *intra* et sensible à la variabilité *inter*.

L'article déjà cité [Ach03] montre qu'une matrice aléatoire, créée à partir de k réalisations indépendantes d'une variable aléatoire, peut être utilisée comme fonction f dans le lemme de JL. Dans l'article [Bar+06], Baraniuk *et al.* reprennent ces travaux et proposent de choisir pour la variable aléatoire : une loi gaussienne centrée de variance $1/n$, une loi de Bernoulli (chaque coefficient de la matrice vaut $\pm 1/\sqrt{n}$ avec la probabilité $1/2$), ou une loi adaptée (chaque coefficient vaut 0 avec la probabilité $2/3$ ou $\pm\sqrt{3/n}$ avec la probabilité $1/6$). Les deux articles [Ach03] et [Bar+06] font reposer ces choix particuliers pour la fonction f du lemme sur le respect d'inégalités de concentration, qui traduisent le fait que la projection aléatoire sur un espace de dimension inférieure est fortement concentrée autour de son espérance mathématique. Ils font également le lien entre ces inégalités, le lemme de JL et la Propriété d'Isométrie Restreinte (*Restricted Isometry Property* ou RIP), aussi appelée principe d'incertitude uniforme, en lien avec l'échantillonnage compressé.

Les limitations du BioHashing

Les schémas de biométrie révocable ne sont pas exempts de vulnérabilités. Un des buts de la protection des données biométriques est d'éviter les usurpations d'identité ou les attaques par rejeu, à partir d'une donnée protégée compromise. Pour lancer ce type d'attaque, il n'est pas nécessaire à un attaquant de reconstruire en totalité la donnée biométrique originale. Il peut suffire en effet de construire une approximation suffisamment proche – appelée pré-image – qui pourra être présentée au système pour obtenir un accès illégitime, comme nous l'avons étudié dans l'article [LCR13].
Comme on l'a vu auparavant dans le paragraphe consacré au lemme de Johnson-Lindenstrauss, la conservation des distances est la propriété-clé de tout schéma de biométrie révocable mettant en jeu une projection sur des matrices aléatoires. Cette même propriété est pourtant à l'origine de vulnérabilités sur la protection des données biométriques : ces vulnérabilités étaient connues dans le domaine de la fouille de données, comme le montre l'article de Liu *et al.* [LGK06]. Dans l'état de l'art, les attaques contre les schémas de biométrie révocable ont été publiés un peu après. Je vais m'appuyer sur quelques articles pour présenter les limitations ou les verrous scientifiques liés au BioHashing et autres schémas de biométrie révocable : [TGN06], [NJ15], [Gom+18], [Don+19b], [Che+19] et [Don+19a].
L'article de Teoh *et al.* [TGN06] propose une variante de l'algorithme original de BioHashing, reposant sur une projection des données biométriques (le visage) sur une séquence de sous-espaces aléatoires. Les auteurs proposent une extension de cette approche à des sous-espaces multiples, appelée *Random Multispace Quantization* (RMQ). Ils soulignent que, théoriquement, une orthogonalité parfaite des matrices de projection est requise pour garantir la préservation des distances. Cependant, elle est difficile à obtenir en pratique. Pour compenser cette difficulté, ils font remarquer que la préservation de la topologie des caractéristiques extraites (*features*) augmente avec la dimension des sous-espaces aléatoires, avec un maximum atteint lorsque la dimension de projection est égale à la dimension de départ : si la dimension de projection est suffisamment grande, les distributions des scores authentiques et imposteurs seront plus séparées que les distributions avant projection. Autrement dit, les variations intra-classes sont préservées, et les variations inter-classes sont renforcées. Ce raisonnement est fondé sur la technique d'extraction des caractéristiques retenue, à savoir l'analyse discriminante de Fisher. Dans cet article, deux types d'attaques sont considérées : la compromission de la donnée biométrique et celle de la clé. La conclusion des auteurs est que la clé et la donnée biométrique jouent un rôle aussi important dans le schéma RMQ. Pour contrer l'attaque par vol de clé (*stolen-token scenario*), ils proposent d'utiliser un algorithme performant pour l'extraction des caractéristiques, sachant que les performances de la reconnaissance biométrique sont proportionnelles à la qualité de l'extraction.

Dans l'article [NJ15], Nandakumar et Jain soulignent que les propriétés de révocabilité et de non-inversibilité ne sont pas garanties intrinsèquement dans les schémas de biométrie révocable, et qu'il est nécessaire de développer des fonctions à sens unique pour rendre le calcul d'une pré-image plus difficile. D'autres attaques sont possibles visant les techniques à base de salage des données, elles sont détaillées dans l'article de Choudhury *et al.* [Cho+18] : attaque par substitution, attaque par multiplicité d'enregistrements, attaque par association (*linkage attack*), attaque FAR, attaque par mascarade, attaque par escalade (*Hill climbing attack*), attaque par écrasement de la décision finale (*overwriting final decision*). Ces attaques exploitent la présence de la donnée biométrique originale, de la clé, de la donnée transformée, ou l'interception de plusieurs données transformées.
Toujours dans l'article [NJ15], Nandakumar et Jain proposent trois raisons principales au fossé existant entre la théorie (notamment en ce qui concerne les propriétés attendues) et la pratique

pour les systèmes de biométrie révocable :

1. La méthode d'extraction des caractéristiques (*features*).
   Elle doit être robuste à la variabilité *intra* : pour cela, les auteurs proposent d'ajouter une étape d'adaptation de ces caractéristiques, pour augmenter la robustesse aux rotations, translations, et autres distorsions non linéaires. Cette étape, selon eux, n'a pas besoin de garantir les propriétés de non-inversibilité ni de révocabilité.
2. Le compromis entre la non-inversibilité et les performances de reconnaissance.
   Il s'agit de la principale limitation présente dans les schémas de protection des données biométriques. Elle est liée au fait que la transformation des données biométriques doit conserver le pouvoir discriminant du gabarit original. Et simultanément, la donnée transformée devrait révéler le moins d'information possible sur la donnée originale. Les auteurs proposent d'étudier la distribution statistique des caractéristiques biométriques pour définir des schémas de protection adaptés. Ce point de vue est également repris par Dong *et al.* dans l'article [Don+19a] à propos du BioHashing : « un BioCode plus long dispose d'une quantité d'information plus importante, et permet ainsi de meilleures performances de reconnaissance. Cependant, plus d'information disponible diminue la complexité des attaques. »
3. La garantie des propriétés de non-associativité et révocabilité.
   Lors de la parution de l'article, en 2015, aucun consensus n'existe dans la littérature pour mesurer ces deux propriétés. Les auteurs de l'article [NJ15] reprennent la proposition faite par Simoens *et al.* [Sim+12] de remplacer l'usage des termes (trop) génériques de sécurité et de respect de la vie privée respectivement par la non-inversibilité et la non-associativité, dans le cadre de la biométrie révocable.

Concernant la non-inversibilité, elle est définie comme la difficulté à obtenir (soit exactement, soit avec une petite marge d'erreur) le gabarit original à partir d'une donnée transformée. Une métrique classique est l'entropie conditionnelle de Shannon, qui permet de mesurer l'incertitude moyenne sur la donnée originale à partir de la donnée protégée. Cette approche convient bien aux cryptosystèmes biométriques. Quant aux schémas reposant sur une transformation des données, souvent la non-inversibilité y est estimée de façon empirique à partir de la complexité de calcul de l'attaque par inversion. Les auteurs citent l'exemple de la courbe *Coverage-Effort* proposée par Nagar *et al.* [NNJ10], qui permet de mesurer le nombre de tentatives (ou *efforts*) requis pour récupérer une fraction (*coverage*) de la donnée originale. Ils soulignent également que cette approche empirique ne permet pas de traiter toutes les stratégies d'inversion des gabarits par des imposteurs.
Dong *et al.* proposent, dans l'article [Don+19a], d'exploiter la propriété de préservation des distances (inhérente aux techniques de biométrie révocable par construction) pour définir une classe d'attaques basées sur la similarité (*similarity-based attacks*). Ces attaques mettent à profit la fuite d'informations provenant de la corrélation des distances entre l'espace des caractéristiques des données originales et celui des données transformées. Autrement dit, ces attaques utilisent les relations de distances entre les gabarits, plutôt que les distances exactes, pour les inverser [DJT19]. Leur but est de générer une approximation (ou pré-image) de la donnée originale à partir d'une donnée transformée et d'une fonction objectif (avec un algorithme génétique dans les références [Don+19a] et [DJT19]), approximation suffisamment proche pour être acceptée par le système. Les auteurs de [Don+19a] se placent dans l'hypothèse de Kerckhoff : l'attaquant a accès à un gabarit transformé et connaît la fonction de transformation ainsi que ses paramètres. Ils testent leur méthodologie d'attaque sur plusieurs schémas de biométrie révocable : le BioHashing , les filtres de Bloom [Gom+16] (que je ne présente pas, car ils sont réservés aux données biométriques binaires, comme l'IrisCode), le hachage par maximisation

d'indice (*Index of Max hashing*) [Jin+18], le hachage spectral multi-dimensionnel non-linéaire (*Non-linear Multi-Dimensional Spectral Hashing*, NMDSH) [Don+19b] et le schéma *Two-factor Protected Minutia Cylinder-Code* (2PMCC) [FMC14]. La figure 3.9 illustre la conclusion des auteurs : il existe un compromis entre la fuite d'information et la précision.

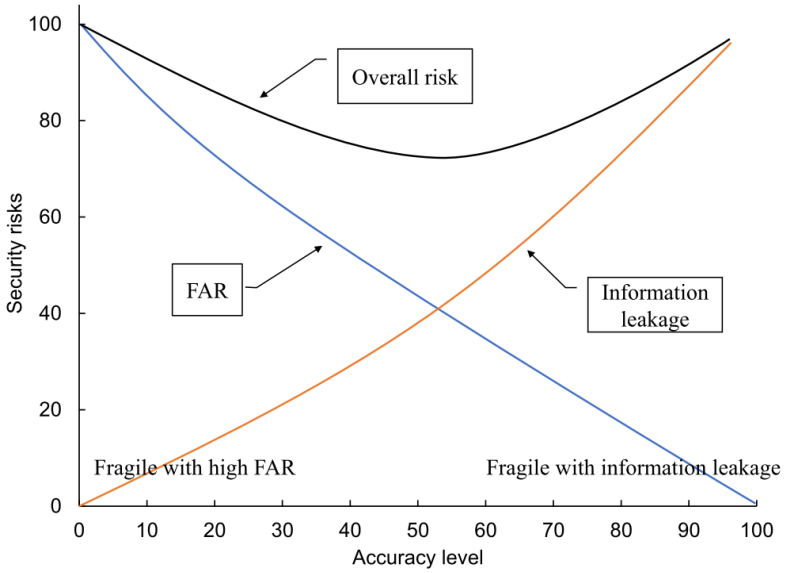


FIGURE 3.9 – Relation entre les risques de sécurité et la précision, extraite de [Don+19a]

Dans l'article [Che+19], Chen *et al.* vont plus loin et démontrent, via un théorème, que lorsque la distance entre deux données biométriques non protégées augmente, la distance entre les données protégées correspondantes augmente également, inévitablement, et ce, quelle que soit la matrice de projection aléatoire. Les auteurs suggèrent, pour éviter que les distances inter entre les données transformées ne dépendent pas des distances intra originales, de choisir une fonction de transformation plus compliquée, non linéaire.

L'article de Dong *et al.* [Don+19b] étudie également la résistance aux attaques basées sur la similarité d'un nouveau schéma de protection révocable, pour le visage, appelé « hachage spectral multi-dimensionnel non-linéaire » (*Non-linear Multi-Dimensional Spectral Hashing*, NMDSH). Même si les caractéristiques de type *deep face* présentent des performances de reconnaissance élevées, elles présentent des vulnérabilités en matière de sécurité et respect de la vie privée. Par exemple, des réseaux neuronaux déconvolutionnels sont capables de reconstruire des images de visages à partir de ces caractéristiques. Les auteurs proposent un nouveau schéma de protection adapté aux caractéristiques de type *deep face*, qui garantit la propriété de non-inversibilité et qui est renforcé pour résister aux attaques basées sur la similarité.

Gomez-Barrero *et al.* proposent explicitement un cadre pour évaluer la non-associativité des données protégées par les schémas de biométrie révocable [Gom+18]. Les auteurs proposent ainsi des métriques et introduisent les notions d'échantillons appariés (*mated samples*) : ce sont deux échantillons (l'un étant une référence stockée et l'autre une requête) provenant du même utilisateur, de la même instance (par exemple, deux empreintes obtenues à partir du même index droit). *A contrario*, deux échantillons sont dits non appariés (*non-mated samples*) s'ils ne proviennent pas de la même instance (par exemple deux empreintes de deux doigts différents). L'article propose une définition de l'associativité dépendante de la méthode utilisée pour réaliser

l'association :
« *Two templates are* fully linkable *if there exists some method to decide that they were extracted, with all certainty, from the same biometric instance. Two templates are* linkable to a certain degree *if there exists some method to decide that it is more likely that they were extracted from the same instance than from different instances.* »
Les auteurs définissent deux fonctions – une globale et une locale – pour mesurer un score de non-associativité entre deux échantillons, appariés ou non. La méthodologie proposée est testée sur des bases de données générées par les auteurs, pour créer des bases d'échantillons appariés et non-appariés.

Au moment de la soutenance de la thèse de Rima Belguechi [Bel15], en 2015, quelques limitations du BioHashing commençaient à émerger dans les différents travaux publiés. Par exemple, le fait que le scénario de vol de la clé était traité en choisissant un seuil de décision différent de celui du système idéal (sans vol de clé ni de biométrie) pour conclure que l'algorithme résistait bien à cette attaque. Ou encore les propriétés de non-associativité ou d'irréversibilité semblaient garanties intrinsèquement par le BioHashing, sans aucune évaluation concrète. Ces thématiques émergentes ont permis à Rima d'apporter de réelles contributions :

- une méthodologie d'évaluation de la sécurité et de la protection des données dans les schémas de biométrie révocable ;
- une amélioration des performances de vérification du BioHashing pour l'empreinte digitale grâce à l'utilisation de descripteurs globaux puis locaux ;
- une architecture de BioHashing sur carte à puce.

J'ai choisi de ne présenter que les deux premières contributions. Les détails de la dernière contribution se trouvent dans l'article [Bel+13].

3.3 Evaluation de la sécurité et du respect de la vie privée dans les systèmes de biométrie révocable

La première contribution de la thèse de Rima que j'ai choisi de présenter porte sur la définition d'une méthode générique d'évaluation des systèmes de biométrie révocable, en particulier le BioHashing. Cette méthode a été développée à partir des travaux de Simoens *et al.* [Sim+12], de la thèse de X. Zhou [Zho12] et du formalisme proposé par Nager *et al.* [NNJ10]. Dans sa thèse, X. Zhou propose quatre étapes pour construire un cadre d'évaluation, appliqué à l'algorithme de *fuzzy commitment* (un cryptosystème biométrique, *cf.* la figure 3.4) :

1. La détermination des objectifs de protection (sécurité, non-associativité, irréversibilité, etc.) ;
2. La détermination du modèle d'attaque (en fonction des capacités d'un imposteur, des informations et ressources disponibles) ;
3. La détermination des métriques d'évaluation ;
4. L'évaluation et l'analyse.

L'article [Sim+12] unifie les objectifs de protection en établissant un ensemble de critères communs applicables à toutes les méthodes de protection de la biométrie, à partir du modèle défini dans la norme ISO/IEC 24745 : la précision du système biométrique ; la dégradation de la précision ; la capacité de traitement ; les besoins de stockage ; la capacité de diversification ; la sécurité et le respect de la vie privée ; la non-inversibilité ; la non-associativité ; la confidentialité et l'intégrité ; la révocabilité ; la capacité de renouvellement ; la séparation des données ; l'indépendance vis-à-vis de la modalité retenue ; l'interopérabilité ; la granularité et la stabilité des performances. Ces critères ne sont pas traduits en termes de métriques : il n'existe donc pas

d'évaluation standardisée des systèmes de biométrie révocable. Les travaux qui s'intéressent aux métriques reposent sur deux approches, une théorique et une pratique. Les métriques théoriques s'appuient globalement sur l'entropie de Shannon pour valider la sécurité des protocoles, ou encore l'entropie conditionnelle et l'information mutuelle. Les métriques pratiques, quant à elles, se concentrent sur la mesure de la complexité des attaques. Dans le chapitre 2 de sa thèse, Rima Belguechi propose une synthèse des différentes attaques connues en 2015 : attaque à zéro effort (qui exploite le FMR) ; attaque par vol de clé ; attaque par association (*linkage attack*) ; estimation de l'inverse ou du pseudo-inverse (appelée aujourd'hui attaque par pré-image, ou par similarité) ; attaque par références multiples ou par écoute ; attaque par force brute ; attaque par vol de la donnée biométrique.
Rima a proposé un cadre pratique, représenté à la figure 3.10, pour l'évaluation des schémas de biométrie révocable, en se concentrant sur le BioHashing, car il est quasi impossible de définir un cadre générique pour toutes les techniques de protection des gabarits biométriques.

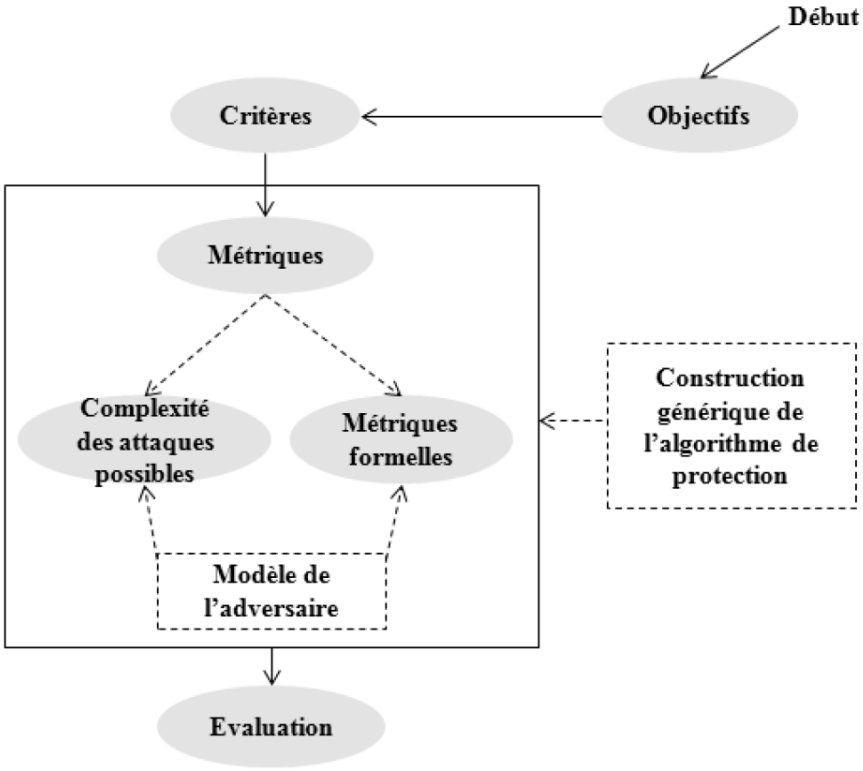


FIGURE 3.10 – Cadre d'évaluation des schémas de biométrie révocable, extrait de [Bel15]

Ce cadre reprend les étapes de la thèse [Zho12] et y ajoute la prise en compte du modèle de l'adversaire, ou de l'attaquant, dont dépendent les métriques considérées. Ce modèle comporte deux types d'adversaires :

1. **L'attaquant malicieux, ou malveillant, ou actif**
   C'est un attaquant actif qui connaît le système. Il a ainsi connaissance de toutes les fonctions, les paramètres et les données générales de l'enrôlement et la vérification. Il peut modifier les informations utilisées pendant la vérification.
2. **L'attaquant honnête mais curieux, ou semi-honnête, ou passif**
   Ce type d'adversaire ne peut pas modifier des informations mais a lui aussi connaissance de toutes les fonctions, les paramètres et les données générées durant l'enrôlement et la vérification.

On adopte les notations définies par Nagar *et al.* dans l'article [NNJ10] :

- b_z, b'_z représentent respectivement, le gabarit et la requête de l'utilisateur z. On garde la même notation b_z pour tous les modèles de référence (quelle que soit la variation intra-classe qui puisse exister entre eux) ;
- on note f la fonction de transformation, qui a comme variables un gabarit et un jeu de paramètres, et n la dimension en sortie de f ;
- K_z, l'ensemble des paramètres de la transformation correspondant à l'utilisateur z ;
- D_O et D_T, les fonctions de distance entre les modèles biométriques dans le domaine original (indice O) et le domaine de la transformation (indice T), respectivement.

Le résultat de la décision du module de vérification du système de biométrie révocable est donné par :

$$R_z = \mathbb{1}_{\{D_T(b_z, b'_z) \leq \epsilon\}} \tag{3.2}$$

où ϵ est le seuil de décision.
Ensuite, Nagar *et al.* définissent les taux d'erreurs avant transformation :

$$FRR_O(\epsilon) = P\left(D_O\left(b_z, b'_z\right) \geq \epsilon\right) \tag{3.3}$$

$$FAR_O(\epsilon) = P\left(D_O\left(b_z, b'_y\right) < \epsilon\right) \tag{3.4}$$

et après transformation :

$$FRR_T(\epsilon) = P\left(D_T\left(f(b_z, K_z), f(b'_z, K_z)\right) \geq \epsilon\right) \tag{3.5}$$

$$FAR_T(\epsilon) = P\left(D_T\left(f(b_z, K_z), f(b'_y, K_y)\right) < \epsilon\right) \tag{3.6}$$

La figure 3.11 présente une vue d'ensemble des métriques proposées dans la thèse de Rima Belguechi. Elles sont classées comme critères de sécurité (colonne de gauche) ou de préservation de la vie privée (colonnes du milieu et de droite). Concernant le profil de l'attaquant, le modèle *malicieux* est retenu pour l'évaluation de la sécurité, tandis que le modèle *honnête mais curieux* (qui peut devenir actif dans certains cas) est retenu pour le respect de la vie privée. J'ai choisi de ne présenter que quelques critères parmi les dix-sept proposés. On notera I_z l'information détenue par l'attaquant au sujet de l'utilisateur z.

- A_3 *: attaque à zéro effort*
  Dans ce scénario, l'attaquant tente d'usurper la véritable identité de l'utilisateur z en représentant ses propres données biométriques b'_y avec des paramètres inconnus K_y. On a donc $I_z = f(b'_y, K_y)$ et la métrique A_3 est définie comme la probabilité de succès de l'attaque, donc le taux de FAR :

$$A_3 = P\left(D_T\left(f(b_z, K_z), f(b'_y, K_y)\right) \leq \epsilon\right) \tag{3.7}$$

- A_5 *: attaque par vol de clé*
  L'attaquant a obtenu la clé de l'utilisateur z, et il essaie différentes valeurs b' pour la requête biométrique, donc on a $I_z = f(b', K_z)$. La métrique A_5 est toujours définie comme le taux de FAR :

$$A_5 = P\left(D_T\left(f(b_z, K_z), f(b', K_z)\right) \leq \epsilon\right) \tag{3.8}$$

- A_{11} *: estimation d'une pseudo-inverse*
  On cherche à mesurer la possibilité de calculer une approximation $\tilde{b}_z$ du gabarit de référence non transformé b_z telle que $\tilde{b}_z$ ne correspond pas à b_z dans le domaine d'origine mais la transformée $f(\tilde{b}_z, K_z)$ correspond à la donnée transformée $f(b_z, K_z)$. Le but de l'attaque est d'optimiser le problème d'inversion lorsque l'estimation de f^{-1} est difficile.

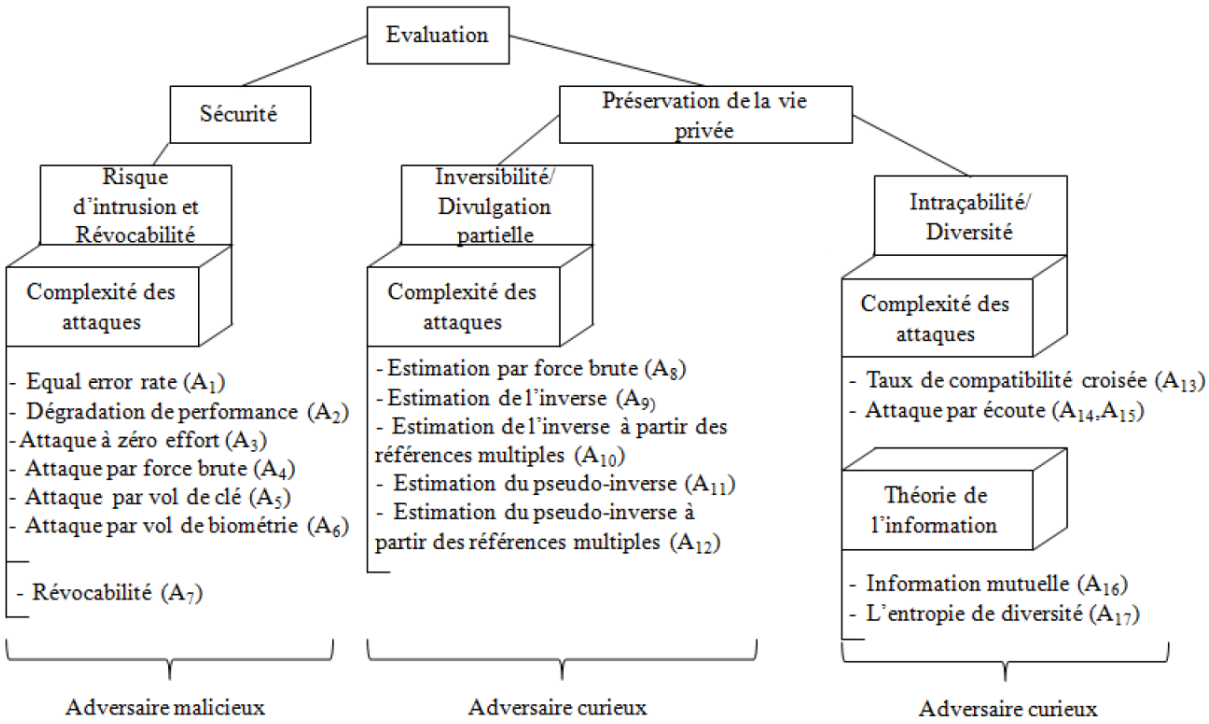


FIGURE 3.11 – Cadre opérationnel pour l'évaluation des transformations révocables [Bel15]

On distingue deux cas : (i) si l'attaquant est honnête mais curieux, A_{11} peut prendre deux valeurs, OF (Optimisation Faisable) et ONF (Optimisation Non Faisable) ; (ii) si l'attaquant devient actif, la métrique est définie par :

$$A_{11} = P\left(D_T\left(f(b_z, K_z), f(\tilde{b}_z, K_z)\right) \leq \epsilon\right) \tag{3.9}$$

- A_{14}, A_{15} *: attaque par écoute*
  On suppose que l'attaquant a intercepté N gabarits transformés ($N = 3$ pour A_{14} et $N = 11$ pour A_{15}) et crée une requête en prédisant la valeur la plus probable pour chaque bit à partir des gabarits obtenus. Les deux métriques A_{14} et A_{15} ont pour valeur le taux de FAR lorsque l'attaquant présente ces requêtes.
- A_{16} *: mesure de la diversité*
  L'information mutuelle I entre les modèles révoqués va permettre de mesurer la capacité de diversité du schéma de biométrie révocable. La métrique est donc définie comme la moyenne de la plus haute valeur de l'information mutuelle, calculée sur tous les utilisateurs :

$$A_{16} = \frac{1}{N}\sum_{z}\sum_{i,j=1}^{M}\max(I\ (f(b_z^i, K_z^i), f(b_z^j, K_z^j))) \tag{3.10}$$

  avec $I(X,Y) = \sum_x\sum_y P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)$, M le nombre de modèles générés pour chaque utilisateur (une valeur $M = 10$ semble réaliste).

Ce cadre d'évaluation sera appliqué dans le paragraphe suivant (plus précisément le sous-paragraphe 3.4.1), pour évaluer et comparer deux schémas de biométrie révocable.

3.4 Amélioration des performances de vérification du BioHashing pour l'empreinte digitale

Dans sa thèse, Rima Belguechi a analysé les performances du BioHashing, en matière de reconnaissance, en fonction du type d'extraction de caractéristiques de l'empreinte digitale. Elle a étudié à la fois des descripteurs globaux (exploitant des attributs de texture) et des descripteurs locaux (les minuties).
Les descripteurs globaux présentent plusieurs avantages. L'utilisation de ces descripteurs aboutit le plus souvent à une représentation de taille fixe qui peut être invariante aux distorsions géométriques. Un descripteur vectoriel est plus facile à intégrer sur un dispositif embarqué car le module de comparaison peut se réduire à un simple calcul de distance entre vecteurs. Les attributs de texture sont plus sécurisés par rapport aux minuties en terme de falsification de l'empreinte car ils ne permettent pas la reconstruction de l'image originale.
Le choix des minuties pour représenter une empreinte est plus efficace au niveau des performances de reconnaissance. Dans sa thèse, Rima Belguechi développe une extension aux gabarits de minuties de l'approche précédente. Elle propose de représenter l'empreinte digitale avec deux descripteurs :

- un descripteur de texture global autour de la minutie pour capturer l'information structurelle avoisinante ;
- un descripteur local

Le défi réside dans le fait qu'un ensemble de minuties n'est pas ordonné, n'est pas de taille fixe, est difficile à extraire si l'image est de mauvaise qualité et surtout cet ensemble de minuties peut permettre de reconstruire l'empreinte originale et ainsi créer une fausse empreinte qui sera acceptée par le système. Par conséquent, il existe peu d'études du BioHashing sur des minuties.

Etude des descripteurs globaux

Rima a testé plusieurs descripteurs de texture : un banc de filtres de Gabor, la méthode LBP (*Local Binary Pattern*), la méthode PBLBP (*Patch Based Local Binary Pattern*), la méthode LRS (Local Relational String). Les tests montrent que la méthode la plus efficace est celle des filtres de Gabor, qui donne le meilleur EER sur les bases FVC2002 [Bel+16].
Une fois ce descripteur de texture sélectionné, le processus représenté à la figure 3.12 est appliqué à l'image pour générer le FingerCode.

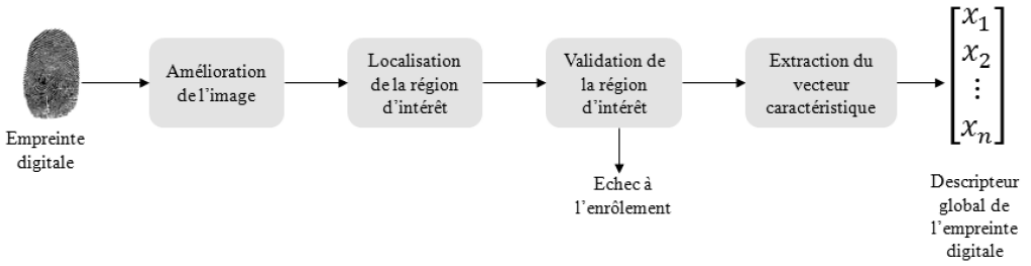


FIGURE 3.12 – Processus d'extraction du descripteur global d'empreintes digitales, extrait de [Bel15]

La première étape d'amélioration de l'image comporte plusieurs pré-traitements (binarisation par exemple).

La deuxième étape est la localisation de la région d'intérêt (ROI, *Region Of Interest*). C'est une étape essentielle pour améliorer les résultats du descripteur de Gabor seul (approche holistique). Elle a comme objectifs :

- sélectionner un point de référence pour prendre en charge les problèmes de translation et/ou de rotation,
- calculer le descripteur à partir d'une sectorisation de l'image (contrairement à l'approche holistique) ce qui permet de mieux gérer l'information contextuelle de l'empreinte.

La ROI peut être circulaire ou carrée, *cf.* la figure 3.13. Elle est validée dans la troisième l'étape.

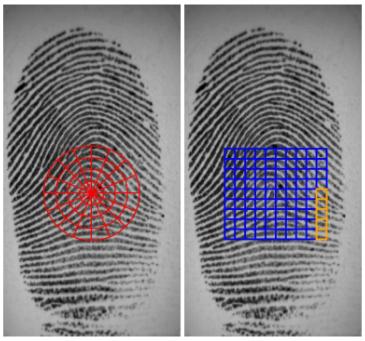


FIGURE 3.13 – Configurations de la région d'intérêt : circulaire (à gauche), carrée (à droite)

Enfin, la quatrième étape est consacrée à l'extraction des caractéristiques pour aboutir au FingerCode : un banc de filtres de Gabor est appliqué sur la région d'intérêt binarisée.
Le problème de rotation est résolu par l'utilisation d'une base d'apprentissage sélectionnée aléatoirement à partir du benchmark de test selon deux approches :

- le gabarit de référence est un vecteur égal à la moyenne des gabarits de test ;
- on fait appel à la théorie bayésienne : le facteur de vraisemblance est utilisé pour calculer la probabilité qu'un individu appartienne ou non à la classe de référence (un utilisateur = une classe représentée par un modèle constitué d'une moyenne et d'une covariance).

Les résultats obtenus sont présentés dans le tableau 3.1. On constate que les différentes étapes du processus proposé ont nettement amélioré les performances, surtout avec la ROI carrée.

Méthode	EER
ROI circulaire	14.17%
ROI carrée	12.77%
Vecteur moyenne (ROI carrée)	10.25%
Facteur de vraisemblance (ROI carrée)	5.14%

Tableau 3.1 – Performances de la vérification d'empreintes par descripteurs globaux

Etude des descripteurs locaux

Pour remédier à l'impossibilité de générer un gabarit de taille fixe comportant toutes les minuties extraites, la stratégie est de protéger chaque minutie avec l'algorithme de BioHashing, et de considérer que le gabarit complet comporte M BioCodes, M étant le nombre de minuties extraites. Rima propose une extension de l'approche précédente pour les minuties. L'empreinte digitale est représentée par deux descripteurs :

1. Un descripteur de texture autour de la minutie pour capturer l'information structurelle avoisinante ;
2. Un descripteur basé minutie qui définit les relations entre chaque minutie et son voisinage local.

Le descripteur de texture sera protégé par l'algorithme de BioHashing et le descripteur basé minutie sera conservé en clair (une analyse de sécurité montre que cela n'engendre aucune vulnérabilité). Le processus est décrit par la figure 3.14 : le gabarit final est composé d'un ensemble de MinuCodes et de K-plets.

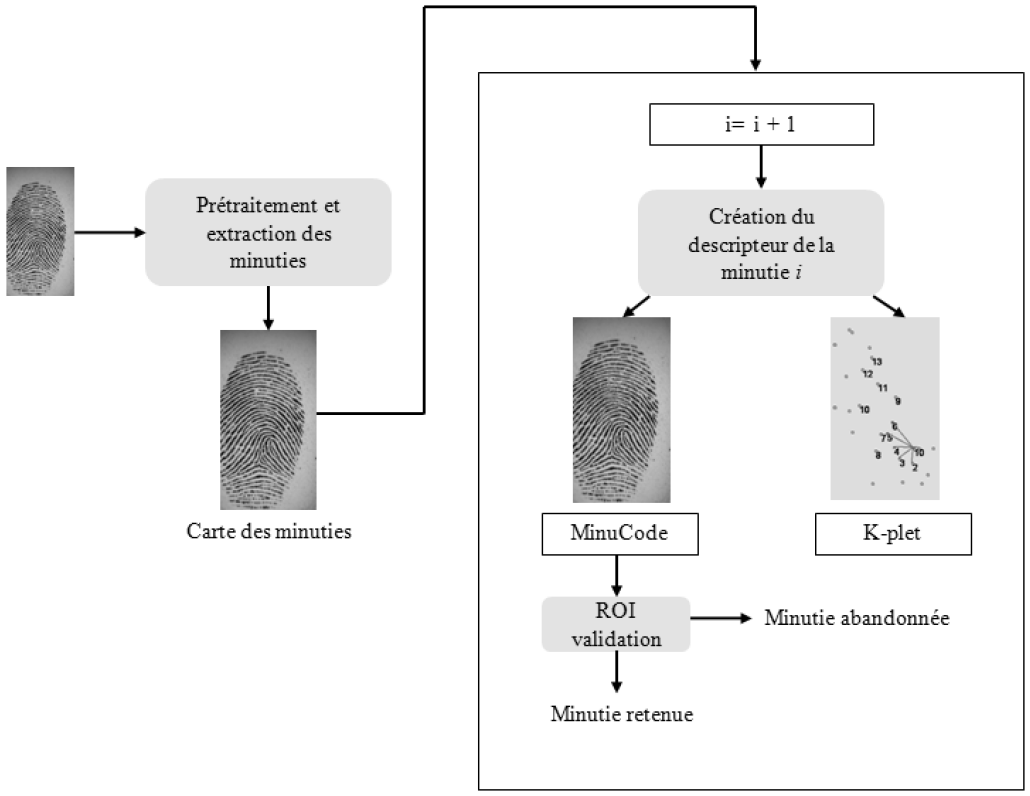


FIGURE 3.14 – Processus d'extraction du descripteur local d'empreintes digitales, extrait de [Bel15]

Autour de chaque minutie, une région d'intérêt circulaire est définie de la même manière que pour les descripteurs globaux (mêmes étapes, *cf.* la figure 3.12). Le résultat de cette étape est un MinuCode, de taille fixe. Même si les performances d'une ROI carrée sont meilleures, une ROI circulaire est privilégiée, pour des raisons d'occupation de l'espace et pour des facilités de gestion des rotations. On rappelle qu'il y a une ROI par minutie.
Pour une minutie donnée, le K-plet associé est l'ensemble des minuties les plus proches et dont les directions relatives (par rapport à la minutie de référence) sont croissantes. Ensuite, les ROI sont validées. Une méthode de correction de l'orientation de référence de l'image d'empreinte est développée dans la thèse pour gérer les problèmes d'alignement.
Pour la vérification, je ne donnerai ici que la philosophie de la technique mise au point par Rima. Intuitivement, deux minuties (l'une appartenant au gabarit de référence, l'autre au gabarit de requête) sont appariées si elles satisfont des contraintes locales et globales :

- la contrainte locale est satisfaite si la distance de Hamming entre leurs MinuCodes protégés est suffisamment petite relativement à un certain seuil ;

- la contrainte globale est satisfaite si la minutie de la requête est géométriquement proche de la minutie de référence, et pour cela, on a recours aux deux K-plets (du gabarit et de la requête).

Les résultats obtenus sont reportés dans le tableau 3.2, qui comporte trois scénarios : le système biométrique sans protection (*Sole Biometric*), le système de biométrie révocable, dans le cas idéal sans vol (*Best Case*) et dans le cas d'un vol de la clé (*Worst Case*). Il faut noter que le seuil de décision retenu est celui du scénario avec vol de clé, pour protéger le système contre les attaques, mais au détriment d'un EER non nul pour les deux autres scénarios.

	Sole Biometric (%)	Best case (%)	Worst case (%)
FVC2002-DB1	1.91	1.75	3.78
FVC2002-DB2	4.56	3.1	6.68
FVC2002-DB3	8.95	4.89	10.87
FVC2002-DB4	10.94	5.38	12.39

Tableau 3.2 – Performances de la vérification d'empreintes par descripteurs locaux

3.4.1 Evaluation des deux approches

La méthodologie développée dans la contribution précédente permet, grâce à l'évaluation des différentes métriques, de comparer les performances des deux systèmes, avec descripteurs globaux et locaux, en matière de sécurité, de respect de la vie privée, et de résistance aux différentes attaques possibles.

Les tableaux 3.3 et 3.4 présentent les résultats des différentes métriques proposées dans la thèse de Rima Belguechi. On peut remarquer qu'aucune attaque n'est effective : quel que soit le seuil de décision, le FAR est à 0%, mis à part l'attaque par vol de clé qui peut réussir dans 13.11% des cas pour les descripteurs globaux, valeur améliorée avec les descripteurs locaux, car $A_5 = 7.16\%$ dans ce cas. On constate que les performances de reconnaissance de la méthode d'extraction des caractéristiques biométriques (pour générer le gabarit) ont une incidence non négligeable sur la résistance aux attaques.

A1	A2	A3	A4	A5	A6	A7	A8	A9
11.40%	1	0%	0%	13.11%	0%	true	5760 bits	NI
A10	A11	A12	A13	A14	A15	A16	A17	
IPP	OF	ONF	0%	0%	0%	0	720 bits	

Tableau 3.3 – Evaluation du schéma de biométrie révocable à base de BioHashing avec les descripteurs globaux de l'empreinte, extrait de [Bel15]

3.5 Perspectives

L'article de Jin *et al.* [Jin+18] souligne que les schémas à base de salage, comme le BioHashing, souffrent d'un écart de performances important si on considère le cas idéal sans attaque et le scénario où la clé est volée (*stolen-token scenario*). Les performances élevées du cas idéal imposent une condition non réaliste, à savoir que la clé doit être gardée secrète en permanence. Ceci devrait être évité dans tous les schémas de protection qui requièrent un ou plusieurs paramètres

A1	A2	A3	A4	A5	A6	A7	A8	A9
6.68%	1	0%	0%	7.16%	0%	true	18432 bits	NI
A10	A11	A12	A13	A14	A15	A16	A17	
IPP	OF	ONF	0%	0%	0%	0	180 bits	

Tableau 3.4 – Evaluation du schéma de biométrie révocable à base de BioHashing avec les descripteurs locaux de l'empreinte, extrait de [Bel15]

indépendants pour garantir la révocabilité des données. Le schéma proposé par les auteurs remédie à ce problème. La méthode de hachage par maximisation de l'indice utilise le lemme de JL localement, ce qui permet de conserver localement les distances, tandis que le BioHashing les conserve globalement. Cependant, un article récent de Ghammam *et al.* [Gha+20] a démontré que cette méthode ne résiste pas aux attaques sur son irréversibilité et sa non-associativité supposées.
Aujourd'hui encore, il est quasi impossible de prouver qu'un schéma de biométrie révocable à base de salage est totalement résistant aux attaques de sécurité et sur la vie privée. L'article de Wang et Baskerville [WB19], propose une vision divergente, qui critique les bénéfices de l'authentification à deux facteurs (par rapport à une authentification à un facteur), avec une analyse originale risques/bénéfices/coûts. Voici un extrait de la conclusion de l'article :
« *There is a definite lack of empirical evidence to support the assumptions that a two-factor authentication scheme delivers superior security compared to a strong single-factor scheme.[...] There is an important need for developing a universal authentication performance evaluation standard. The present constellation of diverse performance evaluation methods and measurements makes the various, existing research findings difficult to comparable.* »
Tous ces éléments laissent donc de nombreuses perspectives, dans des directions variées.

4 Conclusion

Cette thématique symbolise pour moi la réconciliation de deux domaines qui me semblaient incompatibles – la biométrie et la protection de la vie privée. Rima est la première doctorante que j'ai encadrée, sa thèse est un marqueur temporel fort de mon changement d'équipe et de thématique de recherche. Aujourd'hui, la biométrie respectueuse de la vie privée me semble un incontournable de la conception d'un système d'authentification biométrique, si on excepte les systèmes commerciaux où les données sont stockées dans un élément sécurisé. Cette thématique se retrouve également dans mes enseignements, puisque j'interviens sur ces sujets dans le module Biométrie du M2 e-Secure à l'Université de Caen. Je présente également le RGPD dans le module de 3A à l'ENSICAEN sur la Cybersécurité et la gestion des risques. Le contenu de mes cours évolue tous les ans, j'y intègre quelques éléments nouveaux issus de mes recherches. J'ai ainsi proposé une nouvelle partie de cours sur la Protection de la vie privée dans l'un des parcours du nouveau M2.

Voici un extrait d'une interview d'Alain Damasio (écrivain de science-fiction), diffusée sur France Culture en 2019 [12] : « *Ce qu'il se passait en milieu fermé dans la prison où on était hypra surveillés, est devenu la norme d'existence aujourd'hui. Ça ne gêne personne au sens où tant qu'on n'a pas le*

12. https://www.franceculture.fr/societe/le-panoptique-a-lorigine-de-la-societe-de-surveillance

retour de pouvoir sur nos vies, tant que quelqu'un ne s'empare pas de ça pour nous "emmerder", pour nous bloquer, pour nous refuser un job, on ne sent pas qu'on est pris dans cette nasse. Mais le fait est qu'on circule dans des prisons à ciel ouvert en fait. »
Cette réflexion est également présente dans le livre de S. Zuboff [Zub19], qui tente d'expliquer pourquoi on a « laissé » Google, Facebook, Instagram, etc., aspirer nos données et nos comportements : cette révolution numérique est *sans précédent* (formule de l'auteure), par conséquent les utilisateurs n'ont pas de repères, pas de modèles auxquels se raccrocher. On a donc laissé faire au départ, de façon naïve et innocente, jusqu'à ce qu'on ne puisse plus se passer des services proposés, qui simplifient la vie quotidienne, et qu'on ferme les yeux sur le prix à payer.
L'Union Européenne a réagi, sans doute un peu tard, en déployant le RGPD pour protéger les données de ses ressortissants. Ce règlement fait des émules, en Californie par exemple. Cependant, les entreprises s'adaptent très vite aux différentes contraintes : les lignes directrices du Conseil européen de protection des données (EDPB) sur le consentement valide publiées en mai 2020 stipulent que les *cookie walls* ne sont pas un moyen valide pour les sites web d'obtenir le consentement des utilisateurs au traitement de leurs données personnelles et à l'utilisation de cookies. Pourtant, on constate aujourd'hui que certains sites web ont trouvé la parade en quelques semaines en rendant leur consultation payante si les cookies sont refusés par les utilisateurs. Ce comportement présente au moins l'avantage de mettre fin à l'hypocrisie du « tout gratuit » sur internet. Le livre [Zub19] met fin à la croyance classique « si c'est gratuit, c'est toi le produit » : les géants d'internet ne cherchent plus à créer des profils d'utilisateurs, ils recourent aujourd'hui à l'intelligences artificielle pour prédire leurs comportements futurs, et c'est cette prédiction qu'ils revendent à des annonceurs, des services marketing, des publicitaires, etc., qui à leur tour vont essayer d'influencer les comportements des utilisateurs.
Il est donc peut-être temps de renverser le modèle et de replacer l'utilisateur au cœur de la collecte de ses données, en appliquant la transparence, la minimisation, le consentement éclairé, imposés par l'application du RGPD. On peut également aller plus loin en redonnant à l'utilisateur une souveraineté sur ces aspects, une compréhension des enjeux et un réel pouvoir de décision. Tous ces aspects constituent le cœur de mon projet de recherche, développé dans le chapitre suivant.
Pour finir ce chapitre, je voudrais citer G. Bernanos, qui écrivait déjà en 1947 [13] :
« *L'idée qu'un citoyen, qui n'a jamais eu affaire à la Justice de son pays, devrait rester parfaitement libre de dissimuler son identité à qui il lui plaît, pour des motifs dont il est seul juge, ou simplement pour son plaisir, que toute indiscrétion d'un policier sur ce chapitre ne saurait être tolérée sans les raisons les plus graves, cette idée ne vient plus à l'esprit de personne.* »

13. La France contre les robots, Georges Bernanos, 1947

Références du chapitre 4

[Ach03] D. ACHLIOPTAS. « Database-friendly random projections: Johnson-Lindenstrauss with binary coins ». In : *Journal of Computer and System Sciences* (2003).

[Bar+06] R. BARANIUK, M. DAVENPORT, R. DEVORE et M. WAKIN. « The Johnson-Lindenstrauss lemma meets Compressed Sensing ». In : *IEEE Transactions on Information Theory* 52 (2006).

[Bel15] R. BELGUECHI. « Sécurité des systèmes biométriques : révocabilité et protection de la vie privée ». Thèse de doct. Ecole Nationale Supérieure d'Informatique (Alger), 2015.

[Bel+13] R. BELGUECHI, E. CHERRIER, C. ROSENBERGER et S. AIT-AOUDIA. « An integrated framework combining Bio-Hashed minutiae template and PKCS15 compliant card for a better secure management of fingerprint cancelable templates ». In : *Computers & Security* 39 (2013), p. 325-339.

[Bel+16] R. BELGUECHI, A. HAFIANE, E. CHERRIER et C. ROSENBERGER. « Comparative Study on Texture Features for Fingerprint Recognition: Application to The Bio-Hashing Template Protection Scheme ». In : *Journal of Electronic Imaging* 25.1 (2016).

[BCR02] R. M. BOLLE, J. H. CONNELL et N. K. RATHA. « Biometric perils and patches ». In : *Pattern Recognition* (2002). Pattern Recognition in Information Systems.

[Bor14] N. E. BORDENABE. « Measuring Privacy with Distinguishability Metrics:Definitions, Mechanisms and Application to Location Privacy ». Thèse de doct. Ecole Polytechnique, 2014.

[Cav11] A. CAVOUKIAN. *Privacy by Design*. 2011.

[Che+19] Y. CHEN, Y. WO, R. XIE, C. WU et G. HAN. « Deep Secure Quantization: On secure biometric hashing against similarity-based attacks ». In : *Signal Processing* (2019).

[Cho+18] B. CHOUDHURY, P. THEN, B. ISSAC, V. RAMAN et M. K. HALDAR. « A Survey on Biometrics and Cancelable Biometrics Systems ». In : *International Journal of Image and Graphics* (2018).

[Cla11] R. CLARKE. « An evaluation of privacy impact assessment guidance documents ». In : *International Data Privacy Law* (2011).

[Con+05] T. CONNIE, A. TEOH, M. GOH et D. NGO. « PalmHashing: a novel approach for cancelable biometrics ». In : *Information Processing Letters* (2005).

[DFM98] G. I. DAVIDA, Y. FRANKEL et B. J. MATT. « On enabling secure applications through off-line biometric identification ». In : *IEEE Symposium on Security and Privacy*. 1998.

[DJT19] X.-B. DONG, Z. JIN et A. B. J. TEOH. « A Genetic Algorithm Enabled Similarity-Based Attack on Cancellable Biometrics ». In : *IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. 2019.

[Don+19a] X.-B. DONG, Z. JIN, A. B. J. TEOH, M. TISTARELLI et K. WONG. « On the Reliability of Cancelable Biometrics: Revisit the Irreversibility ». In : (2019).

[Don+19b] X. DONG, K. WONG, Z. JIN et J.-L. DUGELAY. « A Cancellable Face Template Scheme Based on Nonlinear Multi-Dimension Spectral Hashing ». In : *2019 7th International Workshop on Biometrics and Forensics (IWBF)*. 2019.

[Eur16] EUROPEAN PARLIAMENT. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.* Available: http://eur-lex.europa.eu/eli/reg/2016/679/oj. 2016.

[FMC14] M. FERRARA, D. MALTONI et R. CAPPELLI. « A two-factor protection scheme for MCC fingerprint templates ». In : *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2014.

[Fli16] C. FLICK. « Informed consent and the Facebook emotional manipulation study ». In : *Research Ethics* (2016).

[Fri84] C. FRIED. « Privacy: A Moral Analysis ». In : *Philosophical Dimensions of Privacy: An Anthology*. Sous la dir. de F. SCHOEMAN. Cambridge University Press, 1984, p. 203-222.

[Gel18] R. GELLERT. « Understanding the notion of risk in the General Data Protection Regulation ». In : *Computer Law & Security Review* (2018).

[Gha+20] L. GHAMMAM, K. KARABINA, P. LACHARME et K. THIRY-ATIGHEHCHI. « A Cryptanalysis of Two Cancelable Biometric Schemes Based on Index-of-Max Hashing ». In : *IEEE Transactions on Information Forensics and Security* (2020).

[GN03] A. GOH et D. C. L. NGO. « Computation of Cryptographic Keys from Face Biometrics ». In : *Communications and Multimedia Security. Advanced Techniques for Network and Data Protection*. Sous la dir. d'A. LIOY et D. MAZZOCCHI. Springer Berlin Heidelberg, 2003.

[Gom+18] M. GOMEZ-BARRERO, J. GALBALLY, C. RATHGEB et C. BUSCH. « General Framework to Evaluate Unlinkability in Biometric Template Protection Systems ». In : *IEEE Transactions on Information Forensics and Security* (2018).

[Gom+16] M. GOMEZ-BARRERO, C. RATHGEB, J. GALBALLY, C. BUSCH et J. FIERREZ. « Unlinkable and irreversible biometric template protection based on bloom filters ». In : *Information Sciences* (2016).

[GS18] A. de GROOT et B. van der SLOOT, éd. *The Handbook of Privacy Studies: An Interdisciplinary Introduction*. Amsterdam University Press, 2018.

[GD99] A. GUPTA et S. DASGUPTA. « An elementary proof of the Johnson-Lindenstrauss Lemma ». In : *Random Structures & Algorithms*. 1999.

[IO13] M. INUMA et A. OTSUKA. « Relations among security metrics for template protection algorithms ». In : *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. 2013.

[JNN13] A. K. JAIN, K. NANDAKUMAR et A. NAGAR. « Fingerprint Template Protection: From Theory to Practice ». In : *Security and Privacy in Biometrics*. Sous la dir. de P. CAMPISI. Springer London, 2013.

[Jin+18] Z. JIN, J. Y. HWANG, Y.-L. LAI, S. KIM et A. B. J. TEOH. « Ranking-Based Locality Sensitive Hashing-Enabled Cancelable Biometrics: Index-of-Max Hashing ». In : *IEEE Transactions on Information Forensics and Security* (2018).

[JL84] W. JOHNSON et J. LINDENSTRAUSS. « Extensions of Lipschitz maps into a Hilbert space ». In : *Contemporary Mathematics* (1984).

[Kin17] E. KINDT. « Having yes, using no? About the new legal regime for biometric data ». In : *Computer Law & Security Review* 34 (déc. 2017). DOI : 10.1016/j.clsr.2017.11.004.

[Koo+16] B.-J. KOOPS, B. NEWELL, T. TIMAN, I. SKORVANEK, T. CHOKREVSKI et M. GALIČ. « A Typology of Privacy ». In : *University of Pennsylvania Journal of International Law* (2016).

[LCR13] P. LACHARME, E. CHERRIER et C. ROSENBERGER. « Preimage Attack on BioHashing ». In : *International Conference on Security and Cryptography (SECRYPT)*. Iceland, 2013.

[17] *Lignes directrices du G29 sur l'analyse d'impact relative à la protection des données - AIPD*. 2017.

[LGK06] K. LIU, C. GIANNELLA et H. KARGUPTA. « An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining ». In : *Knowledge Discovery in Databases*. Sous la dir. de J. FÜRNKRANZ, T. SCHEFFER et M. SPILIOPOULOU. Springer Berlin Heidelberg, 2006.

[Mal+09] D. MALTONI, D. MAIO, A. K. JAIN et S. PRABHAKAR. *Handbook of Fingerprint Recognition*. Springer London, 2009.

[NNJ10] A. NAGAR, K. NANDAKUMAR et A. K. JAIN. « Biometric template transformation: a security analysis ». In : *Electronic Imaging*. 2010.

[NJ15] K. NANDAKUMAR et A. K. JAIN. « Biometric Template Protection: Bridging the performance gap between theory and practice ». In : *IEEE Signal Processing Magazine* (2015).

[RCB01] N. K. RATHA, J. H. CONNELL et R. M. BOLLE. « Enhancing security and privacy in biometrics-based authentication systems ». In : *IBM Systems Journal* (2001).

[Rei95] J. REIMANN. « Driving to the Panopticon: A Philosophical Exploration of the Risks to Privacy Posed by the Highway Technology of the Future ». In : *Santa Clara High Technology Law Journal* 11.1 (1995), p. 27-44.

[Roe05] B. ROESSLER. *The Value of Privacy*. Cambridge: Polity Press, 2005.

[San+19] R. SANCHEZ-REILLO, I. ORTEGA-FERNANDEZ, W. PONCE-HERNANDEZ et H. C. QUIROS-SANDOVA. « How to implement EU data protection regulation for R&D in biometrics ». In : *Computer Standards & Interfaces* (2019).

[SP17] M. SANDHYA et M. V. N. K. PRASAD. « Biometric Template Protection: A Systematic Literature Review of Approaches and Modalities ». In : *Biometric Security and Privacy - Opportunities & Challenges in The Big Data Era*. Sous la dir. de R. JIANG, S. AL-MAADEED, A. BOURIDANE, P. D. CROOKES et A. BEGHDADI. Springer, 2017.

[Sim+12] K. SIMOENS, B. YANG, X. ZHOU, F. BEATO, C. BUSCH, E. M. NEWTON et B. PRENEEL. « Criteria towards metrics for benchmarking template protection algorithms ». In : *2012 5th IAPR International Conference on Biometrics (ICB)*. 2012.

[Sol07] D. J. SOLOVE. « 'I've Got Nothing to Hide' and Other Misunderstandings of Privacy ». In : *San Diego Law Review* (2007).

[Sol21] D. J. SOLOVE. « The Myth of the Privacy Paradox ». In : *George Washington Law Review* (2021).

[TGN06] A. B. J. TEOH, A. GOH et D. C. L. NGO. « Random Multispace Quantization as an Analytic Mechanism for BioHashing of Biometric and Random Identity Inputs ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2006).

[TNG04] A. B. J. TEOH, D. C. L. NGO et A. GOH. « Biohashing: two factor authentication featuring fingerprint data and tokenised random number ». In : *Pattern Recognition* (2004).

[TKL08] A. B. J. TEOH, Y. W. KUAN et S. LEE. « Cancellable biometrics and annotations on BioHash ». In : *Pattern Recognition* (2008).

[WE18] I. WAGNER et D. ECKHOFF. « Technical Privacy Metrics: A Systematic Survey ». In : *ACM Computing Surveys* (2018).

[WB19] P. WANG et R. BASKERVILLE. « The Case for Two-Factor Authentication- Evidence from a Systematic Literature Review ». In : *Twenty-Third Pacific Asia Conference on Information Systems*. 2019.

[WP10] Y. WANG et K. N. PLATANIOTIS. « An Analysis of Random Projection for Changeable and Privacy-Preserving Biometric Verification ». In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* (2010).

[WB90] S. D. WARREN et L. D. BRANDEIS. « The right to privacy ». In : *Harvard Law Review* (1890).

[Zho12] X. ZHOU. « Privacy and security assessment of biometric template protection ». Thèse de doct. Fachbereich Informatik Universitat Darmstadt, Germany, 2012.

[Zub19] S. ZUBOFF. *Age of Surveillance Capitalism : The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs, 2019.

CHAPITRE

4

Projet de recherche

L'imagination est plus importante que le savoir. Le savoir est limité alors que l'imagination englobe le monde entier, stimule le progrès, suscite l'évolution

Albert Einstein

1 Introduction

Ce projet de recherche a mûri depuis deux ans. Je l'ai rédigé en soumettant mon dossier d'HDR à l'École Doctorale MIIS. Depuis, en parallèle de la rédaction de ce manuscrit, j'ai mis en place un certain nombre d'actions qui m'ont permis d'élargir encore le spectre du projet dans sa forme initiale.

- J'ai organisé un Groupe de Travail avec des collègues et quelques partenaires industriels intéressés par la collecte de données comportementales. Notre groupe est constitué d'une dizaine de personnes, nous nous réunissons chaque mois depuis octobre 2020. Le projet a débuté par une journée de *Story Mapping*, un atelier issu des méthodes Agiles, en présence de tous les participants, journée animée par un coach Agile. Cette journée a permis d'aligner les réflexions de tous les participants sur le sujet de la collecte des données personnelles respectueuse de la vie privée.
- J'ai encadré, en binôme avec Sylvain Vernois, ingénieur de recherche, un projet 2A à l'Ensicaen, pour développer une base de l'application de collecte.
- J'encadre actuellement, toujours avec Sylvain Vernois, deux stages d'étudiants de 2A de l'Ensicaen, pour continuer à développer l'outil de collecte.
- Ces trois expériences (Groupe de Travail, projet et stages) m'ont amenée à expérimenter les méthodes Agiles dans un contexte de projet de recherche : ces méthodes sont parfaitement adaptées et pourtant peu appliquées ou étudiées dans la littérature. Ces constats m'incitent à continuer ces expérimentations et à proposer un cadre, un formalisme pour permettre à d'autres équipes de les mettre en place au sein de la recherche.
- J'ai déposé un projet RIN Emergent (Réseaux d'Intérêts Normands) intitulé *Privacy Pre-*

serving Data Collection, P2DC en janvier 2021, qui n'a pas été retenu.
- Cette réflexion pour la constitution du dossier du projet m'a permis d'envisager des pistes de recherche pluridisciplinaires :
  - J'ai entamé une collaboration avec un juriste spécialiste du Droit du numérique : j'ai déposé, avec Thibault Douville, Professeur et Directeur de l'Institut Demolombe, une proposition à l'Appel à Manifestation d'Intérêt *Science Avec et Pour la Société* (AMI-SAPS) de l'ANR, intitulée *La biométrie respectueuse des droits et libertés (BiometricLaw)*. Nous envisageons de définir un sujet de projet pédagogique commun, qui permettrait de faire travailler ensemble des étudiants en informatique de l'Ensicaen et des étudiants en droit du numérique à l'Université de Caen.
  - Je vais contacter des psycho-sociologues pour creuser les aspects d'utilisabilité de l'outil de collecte, la motivation des volontaires au fil du temps, etc.

Mon projet de recherche s'articule en trois temps : un projet à court terme, qui pourra débuter assez rapidement ; un projet à plus long terme, qui nécessite l'achèvement du premier projet ; un projet exploratoire, de recherche fondamentale. On retrouve dans ces différentes propositions les trois aspects déclinés dans la synthèse de mes travaux : la sécurité, l'utilisabilité et la protection des données biométriques. Il est tout à fait fondamental de considérer les trois aspects pour éviter les dérives sécuritaires notamment et préserver les libertés individuelles. Ainsi la sécurité apportée par la biométrie implique une confiance élevée dans l'identité revendiquée par l'utilisateur et permet le déploiement d'applications d'authentification forte pour accéder à de nombreux services numériques. L'utilisabilité permet, quant à elle, une réelle acceptation de la biométrie (visage et empreinte surtout) par le grand public, en complément ou même parfois en remplacement de moyens plus traditionnels d'authentification tels que les codes PIN ou les mots de passe qui encombrent nos mémoires. Enfin, la protection des données biométriques constitue une protection essentielle des libertés individuelles. Plus largement, la protection des données personnelles est mise en lumière depuis un an par l'entrée en vigueur du RGPD dans tous les pays membres de l'Union Européenne, les plaçant de fait au premier rang des pays soucieux de protéger les données de leurs citoyens.
En lien avec les trois chapitres précédents, mes futurs travaux de recherche vont se développer dans trois directions : (i) à court et moyen terme, la conception d'une application sur smartphone afin de constituer une base de données biométriques de qualité (il s'agira de données comportementales) ; (ii) à plus long terme, l'exploitation des données collectées pour approfondir les thèmes déclinés dans la synthèse de mes travaux ; (iii) une recherche plus fondamentale, centrée sur la modélisation et la protection des données biométriques.

2 Projet à court terme : collecte de données

Au cœur de la recherche en biométrie réside l'exploitation de bases de données. Ces bases sont la plupart du temps publiques ou sont constituées spécifiquement pour une étude par l'équipe dans le cadre d'un projet précis. Même s'il existe un certain nombre de bases de données d'empreintes digitales ou de visages accessibles publiquement, c'est loin d'être le cas pour d'autres modalités moins étudiées, notamment pour les modalités comportementales. En outre, les rares données publiques ne sont pas nécessairement de qualité, ni suffisamment diversifiées en ce qui concerne les utilisateurs présents dans la base et ne permettent pas toujours de valider (ni d'infirmer) les hypothèses que l'on souhaite étudier. J'ai donc le projet à court et moyen terme de créer une nouvelle base de données biométriques, données collectées sur smartphone à partir d'une interaction de type schéma de déblocage, auquel les utilisateurs de smartphones sont déjà

habitués. Cette collecte reposera sur une application Android, en suivant quatre étapes :

1. Conception de l'application ;
2. Développement de l'application ;
3. Déploiement de l'application ;
4. Collecte des données.

Différents aspects d'un travail de recherche vont se greffer sur ce projet, aspects que je vais détailler ci-dessous. La création d'une telle application destinée à récolter des données biométriques est plutôt délicate. D'autant plus qu'il s'agit d'une facette assez peu valorisée d'un travail de recherche, et pourtant essentielle. En effet, le but est de recueillir des données de qualité (cela recouvre plusieurs aspects : la qualité des données biométriques elles-mêmes, qui sont un domaine de recherche à part entière ; des données obtenues de façon réaliste), auprès d'un nombre suffisamment important d'utilisateurs.

La conception de l'application

L'application constitue le support de la collecte des données. Se posent donc des questions d'ergonomie, d'*UX-UI Design*, comme je l'ai largement évoqué dans le chapitre 2. Au début de la thèse de Syed Zulkarnain, nous avons lancé une campagne d'acquisition de données de dynamique de frappe au clavier (sur ordinateur). L'élaboration de *scénarios d'acquisition* et le choix des *paramètres* constituent le socle de la collecte, un passage obligé qui conditionne la réussite de l'exploitation future des données.

Les *scénarios d'acquisition* vont poursuivre deux objectifs :

1. Une collecte des données dans la durée.
   Un des aspects intéressants à étudier est l'évolution (ou la dérive) des données comportementales au cours du temps. Chaque utilisateur devra définir son propre schéma, et le reproduire à chaque session. Le but est d'améliorer l'utilisabilité, et de développer de nouvelles stratégies de mise à jour de modèle biométrique.
2. Un second scénario orienté vers les attaques.
   Le but est d'étudier puis de renforcer la sécurité de l'authentification à base de biométrie comportementale. Pour cela, les utilisateurs devront reproduire plusieurs schémas imposés, de complexité croissante. Des performances d'attaques pourront être calculées en fonction de la complexité du schéma.

La détermination des *paramètres de la collecte* doit être cohérente avec le but de l'étude. J'ai déjà organisé une réunion au sein de mon équipe de recherche pour rassembler les conseils de tous les membres qui ont déjà réalisé une collecte de données et ainsi capitaliser nos expériences passées. Ces paramètres sont nombreux et assez diversifiés :

- le rythme et le nombre de sessions, le délai entre deux sessions ;
- le déroulement d'une session (nombre de captures, schéma personnel ou imposé en fonction du scénario, etc.) ;
- le recrutement de volontaires motivés : via des informations précises, une sensibilisation par rapport aux enjeux de l'étude et des travaux de recherche futurs, afin d'obtenir le consentement éclairé des utilisateurs, en accord avec le RGPD ; mais aussi via la création d'un esprit de groupe ;
- la récolte d'informations pertinentes pour des études futures (par exemple l'âge, le genre, la marque du smartphone…) : en accord avec le RGPD, la collecte doit prendre en

compte notamment les aspects de *proportionnalité*, de *minimisation* et de *finalité*, aspects qui peuvent être assouplis dans le cadre de données de recherche.

Le développement de l'application

Cette étape se fera en collaboration avec Sylvain Vernois, ingénieur de recherche de l'équipe SAFE (l'équipe Monétique & Biométrie a changé de nom depuis janvier 2021). Il a déjà développé une application pour récolter des données biométriques liées à un schéma de déblocage d'un smartphone. Cette application a servi à récolter des données au sein de notre équipe, sur un nombre réduit de volontaires. Nous avons déjà discuté au sujet d'une expérimentation de plus grande envergure. L'architecture implantée étant modulaire, l'application existante pourra être complétée à la fois pour répondre aux besoins de l'étude que j'envisage et pour fixer les différents paramètres détaillés dans le point précédent.

Le déploiement de l'application

Du point de vue des utilisateurs volontaires pour participer à la collecte de données, le déploiement fera suite à la première étape de prise d'information et de sensibilisation.
Pour le premier scénario, il est important que ces deux étapes soient décorrélées dans le temps et en terme d'affichage, car elles ont des finalités tout à fait distinctes. J'ai déjà testé l'envoi d'un lien de téléchargement, qui n'a pas donné de bons résultats en terme de recrutement de volontaires : je suis donc persuadée qu'un accompagnement personnalisé est réellement indispensable lors de cette étape également. Certes, cela nécessite un investissement important en temps, mais cela influe directement sur l'implication des utilisateurs sur toute la durée de l'expérimentation. Une autre option pourrait être de recourir à une société dont le métier est le recrutement de volontaires, qui reçoivent une rémunération. Même si ce type de prestation garantit un échantillon représentatif de la population, le coût reste rédhibitoire pour le moment, en dehors de toute nouvelle demande de financement de projet.
Parallèlement, j'envisage d'inclure le second scénario à mon module d'enseignement de Sécurité des Systèmes d'Information - Cybersécurité : les étudiants seront motivés pour mener les attaques sur les schémas et ils auront une vision critique constructive de la sécurité apportée par le système d'authentification lié à l'application. Une étude de type PIA (Privacy Impact Assessment), détaillée sur le site de la CNIL, pourra ensuite être conduite pour évaluer les risques sur la vie privée.

La collecte des données

L'étape suivante de collecte des données permet de s'assurer que les volontaires respectent le protocole fixé, que les données sont capturées dans les conditions prévues. La pédagogie vient donc s'ajouter à la simplicité d'usage garantie par les phases de conception et de développement. De ce respect du protocole dépend la qualité des données récoltées. De nouveau, un accompagnement des volontaires semble indispensable, au moins lors des premières sessions, puis régulièrement si l'expérience dure longtemps et comporte de nombreuses sessions.
A la fin de la collecte, il faut prévoir une phase de retour d'expérience, pour améliorer les futures collectes en regroupant les ressentis des utilisateurs, leurs remarques, leurs critiques et conserver ce qui a bien fonctionné. Pour développer leur confiance dans l'outil, on pourrait recueillir leur avis en ce qui concerne le niveau de sécurité, la protection de leur vie privée et de leurs données, en fonction des paramètres de chaque collecte.

Ce qui précède correspond à la construction de mon projet avant la crise sanitaire du COVID-19. Les conditions de la future collecte ont bien évidemment évolué : elle se fera sans doute à distance. Il s'agit donc désormais de renforcer l'autonomie de l'utilisateur face à l'outil de collecte. C'est ce sur quoi nous travaillons dans le cadre des stages de 2A, en suivant une méthodologie Agile, avec un regard important sur l'*UX-UI Design*.
Tout au long de cette expérimentation qui met en jeu des données biométriques, le RGPD doit être appliqué. Cette réflexion est tout à fait en lien avec mes missions en tant que RIL (Relais Informatique et Liberté) du laboratoire.

Multiplication des collectes

La réflexion que j'ai menée lors de la rédaction de mon projet RIN m'a amenée à faire évoluer l'outil proposé, en multipliant les collectes. L'idée est de partir d'une hypothèse de recherche à tester, de lancer une nouvelle collecte qui ne capturera que les éléments strictement indispensables à la validation de l'hypothèse. Les objectifs pourraient se décliner comme suit.

- Les utilisateurs devront être éclairés, informés (sur le traitement futur de leurs données), rester maîtres de leurs données (ils peuvent refuser la collecte le cas échéant). Ils donneront leur avis sur leur ressenti par rapport au triptyque : sécurité, utilisabilité, protection de leur vie privée.
- Un minimum de données sera collecté : il s'agit d'extraire uniquement les informations nécessaires pour tester une hypothèse de recherche.
- Les collectes seront nombreuses : grâce à un outil évolutif (d'où une préférence pour une méthodologie agile), dynamique (on lancera une nouvelle collecte quand un nouveau besoin émergera du côté recherche), adaptatif (qui s'adaptera aux pistes de recherche, aux retours des utilisateurs), interactif (en mettant au point des éléments de pédagogie pour sensibiliser l'utilisateur), et surtout convivial (grâce aux principes de l'*UX-UI Design*).

3 Projet à plus long terme : exploitation des données

Lorsque la phase de collecte sera achevée, l'exploitation des données pourra commencer, que je prévois mener sur un temps plus long. La qualité de la base de données ainsi constituée est essentielle pour pouvoir tester plusieurs pistes de recherche. Je vais décliner ces pistes selon les trois axes présents dans la synthèse de mes travaux, à savoir la sécurité, l'utilisabilité et la protection des données biométriques.

Sécurité

Les données comportementales sont moins performantes que les modalités physiologiques. Des pistes d'amélioration ont été proposées dans le chapitre 1 : on peut y remédier d'une part en améliorant les techniques de mise à jour de modèle, d'autre part en travaillant sur l'extraction de caractéristiques à partir des données brutes. Les résultats obtenus pour la mise à jour de modèle appliquée à la dynamique de frappe au clavier dans la thèse d'Abir Mhenni (notamment l'approche exploitant le zoo de Doddington) pourront être étendus à une autre modalité comportementale. Par ailleurs je propose d'exploiter des outils de traitement du signal : application de techniques avancées de reconnaissance du locuteur (i-vectors, e-vectors, représentations parcimonieuses) à une modalité comportementale, utilisation de caractéristiques fréquentielles pour modéliser le comportement d'un utilisateur.
Par la suite, d'autres modalités caractéristiques du comportement de l'utilisateur pourront être

étudiées à partir des interactions entre l'utilisateur et son smartphone et d'autres collectes de données pourront être envisagées.

Utilisabilité

Les retours d'expérience à la fin de la collecte des données constitueront un réel apport pour étudier cet aspect de la biométrie rarement traité dans la littérature. Comme on l'a vu dans le chapitre 2, l'utilisabilité représente pourtant une des explications-clés de l'essor de la biométrie en tant que facteur d'authentification forte. Une réflexion doit permettre d'envisager de nouvelles applications : même si la biométrie est un domaine d'application, elle n'est jamais une finalité, elle est toujours impliquée dans un scénario d'authentification. Ainsi une application Android respectant les principes de l'*UX-UI Design* et performante pourra être valorisée dans différents contextes : dans un contexte médical ou une situation de handicap, ou pour le paiement électronique, au sein d'une architecture sécurisée et respectueuse de la vie privée. L'intérêt de recourir à la biométrie comportementale est d'autant plus grand que son utilisabilité est intrinsèque. Cette piste cependant doit être une résultante du point précédent : l'intégration de l'application développée dans un contexte réel plus global est conditionnée par un niveau de sécurité suffisant.

Protection des données biométriques

La biométrie révocable a été principalement appliquée à des modalités physiologiques (visage, empreinte digitale, iris), comme on l'a vu dans le chapitre 3. Peu de travaux dans ce domaine ont été publiés concernant des modalités comportementales. Il semble donc pertinent d'étudier comment on peut adapter au mieux le BioHashing à des données comportementales : pour pouvoir supporter des scénarios d'attaques sur l'un des deux facteurs (biométrie ou secret), les performances du système biométrique sous-jacent (c'est-à-dire non révocable) doivent être les meilleures possibles. De nouveau, on est ramené à améliorer les performances du système d'authentification à base de biométrie comportementale. Il faut également trouver la modélisation adéquate des données, pour exploiter au mieux la puissance de l'algorithme de BioHashing. Sachant que des filtres de Gabor donnent de bons résultats sur les empreintes digitales, une modélisation fréquentielle serait sans doute une bonne piste.

On garantit ainsi une logique de transparence, de proportionnalité et de finalité, imposées par le RGPD, assortie d'une sensibilisation de l'utilisateur aux problématiques liées à la collecte de ses données comportementales.

4 Projet de recherche fondamentale : modélisation et protection des données biométriques

Si les pistes évoquées dans les points précédents peuvent être envisagées à court ou moyen terme après la constitution de la base de données comportementales, mon projet de recherche s'inscrit également dans un temps plus long. Les idées présentées maintenant viennent d'une prise de recul sur les aspects plus fondamentaux que j'avais mis un peu de côté depuis mon changement d'équipe en 2011. Elles s'orientent dans deux directions : la modélisation des données de biométrie comportementale et la protection des données biométriques.

Modélisation des données biométriques comportementales

Définition d'un multimodèle biométrique

A travers les travaux de la thèse de Syed Zulkarnain (biométrie douce) et surtout ceux d'Abir Mhenni (mise à jour de modèles biométriques), il semble que les performances augmentent lorsque le système peut être personnalisé, c'est-à-dire lorsqu'on prend en compte les particularités de chaque utilisateur (par exemple avec un seuil de décision ou un protocole de mise à jour du modèle personnalisés). Je propose de pousser plus loin cette personnalisation, en combinant la biométrie douce et la mise à jour de modèle. Une partie de mes travaux de thèse a été consacrée aux multimodèles : il s'agit de modéliser la dynamique d'un système comme une combinaison linéaire de plusieurs sous-modèles établis autour de différents points de fonctionnement. Pour les données de biométrie comportementale, la biométrie douce pourrait permettre d'identifier quel est l'état d'esprit de l'utilisateur (fatigué, en colère, de bonne humeur, etc.), puis de créer, dans un premier temps, un modèle de base associé à cet état d'esprit. Dans un second temps, on pourrait considérer qu'à chaque instant, l'état d'esprit de l'utilisateur est une combinaison (linéaire pour commencer) de ses différents états d'esprit. Par conséquent, son comportement serait une combinaison de ses différents modèles de base. Pour un utilisateur donné, on peut supposer que les coefficients de la combinaison linéaire restent dans un intervalle restreint, permettant d'authentifier l'utilisateur. Une modélisation plus fine pourrait être apportée dans un deuxième temps par une collaboration avec des neuropsychologues par exemple.

Une représentation unifiée pour les modalités comportementales

Les données de biométrie comportementale présentent un point commun que j'ai déjà mentionné : elles souffrent d'une grande variabilité *intra-utilisateur*. Le défi de leur modélisation consiste donc à extraire des caractéristiques les moins sensibles à cette variabilité *intra*. Cette variabilité a deux causes principales : (i) notre comportement n'est pas figé, il peut dépendre de notre état d'esprit, notre état de fatigue, etc (*cf.* le point précédent) ; (ii) nous apprenons continuellement, ce qui se traduit par une dérive lente du modèle actualisé par rapport au modèle stocké comme référence.
Même si les données biométriques ne peuvent être décrites comme la réalisation d'un système dynamique, il n'empêche que cette dérive pourrait être vue comme une dynamique sous-jacente très lente. Les techniques classiques de modélisation numérique (filtres de Kalman, modèles AR, ARMA, ARMAX, chaînes de Markov, etc.) ne conviennent pas ici. Je propose d'utiliser d'autres techniques, appartenant au domaine de la modélisation fonctionnelle des données : un jeu de données fonctionnelles constitue, par définition, un ensemble d'échantillons correspondant aux observations d'un objet aléatoire continu. Il semble assez réaliste de considérer que notre comportement dépend d'une infinité de paramètres, et que les courbes obtenues, par exemple, à la suite de frappes de textes au clavier, ou de tracés de schémas sur smartphone, sont liées à un modèle plus global. Cela permettrait d'expliquer ces petites variations *intra* grâce au cadre théorique de la modélisation fonctionnelle.

Protection des données biométriques

Selon le RGPD (article 33), *les données biométriques sont des données personnelles sensibles*, qui doivent donc faire l'objet d'une protection particulière. Comme je l'ai mentionné dans la synthèse de mes travaux de recherche, l'algorithme de BioHashing, au cœur de la thèse de Rima

Belguechi, est particulièrement efficace lorsqu'aucune attaque ne menace le système. Mais certains articles de l'état de l'art ont montré que différentes implémentations du BioHashing (ou ses dérivés) peuvent être attaquées. Je souhaite donc revenir à la base de l'algorithme et travailler sur le lemme de Johnson-Lindenstrauss qui en est le pivot. D'autres articles établissent des liens entre ce lemme et la théorie de l'échantillonnage compressé, via des projections aléatoires entre dimensions finies. Comme mentionné précédemment, la biométrie révocable a été peu appliquée à des modalités comportementales, peut-être en raison de leur caractère moins sensible que les modalités physiologiques (empreinte digitale, visage, iris...). Cependant, en cas de compromission, il ne paraît pas évident de modifier sa façon de taper au clavier, sa façon de parler, sa façon de signer ou de marcher. Étant donné l'appétence des industriels pour la biométrie comportementale qui leur permet de concevoir des architectures d'authentification continue et/ou transparente, il me semble indispensable de protéger également ce type de données. J'envisage donc de mener une étude approfondie du mécanisme au cœur du lemme de Johnson-Lindenstrauss à la fois sur le plan théorique (les articles sur le BioHashing ou équivalent vérifient rarement les hypothèses de ce lemme fondamental) et sur le plan numérique, grâce à la base de données collectée.