Guerre cognitive et influence des décisionnaires : approche méthodologique pour la conception d'un système d'aide au développement de la conscience de situation et à la prise de décision

Marie Morelle-Gerritsen

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DES SCIENCES PHYSIQUES ET DE L'INGÉNIEUR

SPÉCIALITÉ : AUTOMATIQUE, PRODUCTIQUE, SIGNAL ET IMAGE, INGENIERIE COGNITIQUE

Par Marie MORELLE-GERRITSEN

Guerre cognitive et influence des décisionnaires : approche méthodologique pour la conception d'un système d'aide au développement de la conscience de situation et à la prise de décision

Sous la direction de : Jean-Marc ANDRÉ

Soutenue le 26 septembre 2025

Membres du jury :

M. LE BLANC Benoît	Professeur des universités	Bordeaux INP – ENSC	Président du jury
M. VALETTE Mathieu	Professeur des universités	INALCO	Rapporteur
M. CHAUDRON Laurent	Docteur, HDR	CREA	Rapporteur
Mme COLLOMB Cléo	Maîtresse de conférences	Univ. Paris-Saclay	Examinatrice
Mme LEHMANS Anne	Professeure des universités	Univ. Bordeaux, INSPE, lab. IMS	Examinatrice
M. CEGARRA Julien	Professeur des universités	INU Champollion	Examinateur
M. ANDRÉ Jean-Marc	Professeur des universités	Bordeaux INP – ENSC, lab. IMS	Directeur de thèse

Membres invités :

M. MARION Damien	THALES LAS France	Invité
M. GARDINETTI Emmanuel	Agence de l'Innovation de Défense	Invité

Guerre cognitive et influence des décisionnaires : approche méthodologique pour la conception d'un système d'aide au développement de la conscience de situation et à la prise de décision

Résumé

La guerre cognitive vise à influencer, modifier et orienter la pensée humaine ainsi que le traitement de l'information. Elle utilise divers moyens technologiques et numériques, tels que les réseaux sociaux, la transmission d'informations, les cyberattaques et les Nanotechnologies, Biotechnologies, technologies de l'Information et sciences Cognitives (NBIC), ainsi que des combinaisons de ces moyens. La guerre cognitive cible la cognition et l'esprit humain. Elle s'attaque aux représentations mentales et aux processus cognitifs pour semer le doute, empêcher ou influencer des décisions et saper la volonté de l'adversaire. Elle peut aussi manipuler la façon dont les individus perçoivent et interprètent les informations sans qu'ils en soient conscients. L'objectif est d'obtenir un avantage tactique ou stratégique, par la déstabilisation, l'altération de la confiance, l'obtention d'informations ou de décisions souhaitées, l'inhibition de comportements attendus ou la manipulation de l'opinion. Comme le souligne l'ancien chef d'état-major des armées, le général Thierry Burkhard, il s'agit de « gagner la guerre avant la guerre ». L'actualité récente et moins récente met en évidence la nécessité de prendre conscience de la réalité et de l'importance des actions de guerre cognitive et de manipulation de l'information et de la décision, afin de s'en prémunir en étant proactif face aux agents d'influence qui utilisent ces tactiques. Ainsi, nous cherchons à contribuer à cette prise de conscience, en apportant des éléments conceptuels de compréhension, des exemples et des pistes de solutions.

L'objectif de cette thèse est donc de poser les fondements d'un système d'aide à la prise de décision qui propose une détection des actions de guerre cognitive, des contre-mesures et des actions offensives afin d'influencer la situation. Pour répondre à ces exigences, nous avons structuré notre travail en cinq étapes principales :

1) Une approche conceptuelle de la guerre cognitive : définitions, cibles et acteurs, outils, exemples et enjeux.
2) Une étude des biais et influences de la décision et des méthodes pour s'en protéger, notamment avec des systèmes d'aide à la décision.
3) Une expérimentation visant à déterminer les facteurs d'influence de la décision, et plus particulièrement ceux qui rendent un texte plus crédible et lui confèrent donc un potentiel d'influence plus important. Après une étude de la littérature, qui nous a permis de lister un grand nombre de facteurs de crédibilité d'un texte, puis un travail de réduction des facteurs, nous avons mené un tri de cartes en ligne sur les 12 facteurs finaux sélectionnés afin de déterminer lesquels confèrent le plus de crédibilité à un texte.
4) L'élaboration d'un système pour la détection en temps réel et l'influence des stratégies. Cette étape fait l'objet d'une seconde expérimentation, menée avec pour support un jeu de société semi-collaboratif. Nous avons utilisé le principe d'arbre de détection d'informations critiques du système ANTICIPE développé par THALES, pour détecter la stratégie des joueurs en temps réel et proposer des solutions pour influencer leur stratégie. Cette étape se base sur les acquis de la première expérimentation. À partir des résultats de ces différentes étapes, nous proposons une méthodologie pour adapter le système ANTICIPE afin d'en faire un outil d'aide à la décision dans un contexte de guerre cognitive.

5) Une étape de détermination des critères de qualification des cibles de guerre cognitive : nous proposons une liste de 5 critères à étudier afin de déterminer si l'individu ciblé est pertinent, notamment en fonction de son potentiel d'influence sur la décision visée et de ses vulnérabilités.

Les résultats de cette thèse en ingénierie cognitique permettent de poser les bases pour la construction d'un système technologique qui contribue à la prise de conscience de la situation dans un contexte de guerre cognitive et assiste son utilisateur dans la sélection des cibles pertinentes et des solutions pour une réponse cognitive ciblée.

Mots-clés : Conception système cognitif – guerre cognitive – prise de décision – gestion de crise – biais cognitifs

Cognitive Warfare and Influence on Decision-Makers: A Methodological Approach to Designing a System for Supporting Situational Awareness and Decision-Making

Abstract

Cognitive warfare aims to influence, modify and direct human thought and information processing. To achieve this, it employs various technological and digital means, such as social networks, information dissemination, cyber-attacks and Nanotechnology, Biotechnology, Information science and Cognitive science (NBIC) technologies, as well as combinations of these methods. Cognitive warfare targets the human cognition and mind. It targets mental representations and cognitive processes to sow doubt, prevent or influence decisions, undermine the adversary's will, or manipulate how people perceive and interpret information without being aware of it. The goal is to gain a tactical or strategic advantage through destabilization, erosion of trust, acquisition of desired information or decisions, inhibition of expected behaviors, or manipulation of public opinion. As emphasized by the French Chief of the Defense Staff, General Thierry Burkhard, it is about "winning the war before the war". Recent and less recent events highlight the importance of recognizing the reality and significance of cognitive warfare actions and the manipulation of information and decision-making, in order to protect against them by being proactive against agents of influence who employ those tactics. Thus, through this work, we seek to contribute to this awareness by providing conceptual elements of understanding, examples, and potential solutions.

The objective of this thesis is to lay the foundations for a decision support system that offers detection of cognitive warfare actions, countermeasures and offensive strategies to correct the situation. To meet these requirements, we structured our work into five main steps:

1) A conceptual approach to cognitive warfare: definitions, targets and actors, tools, examples and challenges.
2) A study of biases and influences on decision-making and methods to protect against them, particularly with decision support systems.
3) An experiment aimed at determining the factors influencing decision-making, particularly those that make a text more credible and thus give it a greater potential for influence. After reviewing the literature, which allowed us to list a large number of text credibility factors, and then reducing these

factors, we conducted an online card sorting exercise on the 12 final selected factors to determine which ones confer the most credibility to a text.

4) The development of a system for real-time detection and influence of strategies. This step is the subject of a second experiment, conducted using a strategy game. We used the decision tree principle of the ANTICIPE system developed by THALES to detect players' strategies in real-time and propose solutions to steer their strategy. This step builds on the findings of the first experiment. Based on the results of these different steps, we propose a methodology to adapt the ANTICIPE system to make it a decision support tool in the context of cognitive warfare.
5) A step to determine the criteria for qualifying cognitive warfare targets: we propose a list of 5 criteria to study in order to determine if the targeted individual is relevant, particularly based on their potential influence on the targeted decision and their vulnerabilities.

The results of this thesis in cognitive engineering lay the foundation for building a technological system that contributes to situational awareness in a cognitive warfare context and assists its user in selecting relevant targets and solutions for a targeted cognitive response.

Keywords: Designing cognitive system – cognitive warfare – decision making – crisis management – cognitive biases

IMS, groupe cognitique, équipe Cognitique et Ingénierie Humaine

Université de Bordeaux, IMS - CNRS UMR 5218, Talence, France

Remerciements

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce travail de thèse.

Tout d'abord, je remercie les membres du jury présents lors de ma soutenance de thèse pour la qualité et la pertinence de leurs remarques, en particulier les rapporteurs Laurent Chaudron et Mathieu Valette.

Je remercie chaleureusement mon directeur de thèse, Jean-Marc André, ainsi que mon encadrant chez THALES, Damien Marion, pour la qualité de leur accompagnement, leur disponibilité, leurs conseils avisés et leur soutien tout au long de cette thèse.

Je remercie également THALES pour le financement de cette recherche, et plus particulièrement à Erwan Boulain et Olivier Vivien pour leurs conseils et leur soutien en tant que managers.

Je suis reconnaissante envers l'Agence de l'Innovation de Défense pour le co-financement de cette thèse, et tout particulièrement envers Emmanuel Gardinetti, Guillaume Chillet et Baptiste Prébot pour leur suivi attentif et leurs remarques pertinentes, avec une mention spéciale à Guillaume Chillet pour son aide précieuse dans le recrutement de participants pour la première expérimentation.

Un grand merci à Julien Cegarra pour ses précieux conseils et son regard critique sur mes travaux, qui m'ont permis de les enrichir et d'approfondir ma réflexion.

Je souhaite également remercier Maxime Poret et Titouan André-Roure pour le développement des outils de la deuxième expérimentation et leur aide indispensable lors des passations.

Mes remerciements s'adressent aussi à Hélène Unrein et Théodore Letouzé pour leur contribution à la génération d'idées et à l'amélioration des protocoles d'expérimentation.

Je remercie François Demontoux et Laurent Chaudron, membres de mon comité de suivi de thèse, pour leur suivi annuel attentif et bienveillant et leurs excellents conseils.

Je tiens à exprimer ma gratitude envers Bernard Claverie pour les nombreux articles scientifiques fournis, qui ont nourri ma revue de littérature, ainsi que pour son travail inestimable sur le sujet de la guerre cognitive.

Un remerciement particulier à Juliette Mattioli (THALES) pour son aide déterminante dans la diffusion de l'expérimentation 2, qui m'a permis d'atteindre un nombre suffisant de participants.

Je remercie aussi le Général Gilles Desclaux ainsi que Benoît Lamirault pour leurs contributions concernant le modèle des cibles de guerre cognitive.

J'adresse mes remerciements à Maximilien Lorans, qui a effectué un stage dans le cadre de cette thèse, ainsi qu'aux étudiants en projet de fin d'études, Julien Debidour Lazzarini et Lucie Della-Negra, pour la qualité de leur travail et leur implication dans le projet.

Je n'oublie pas l'ensemble des participants aux expérimentations pour le temps qu'ils m'ont consacré : sans eux, ce travail n'aurait pu aboutir. De même, je tiens à exprimer mes plus sincères remerciements à toutes les personnes, collègues, chercheurs, famille, que je n'ai pas citées ici mais qui ont contribué à l'avancée de ces travaux de par nos échanges.

Enfin, un immense merci à mon mari Alexandre Gerritsen pour son soutien indéfectible, sa patience, sa compréhension et son aide précieuse, y compris dans les moments les plus intenses de cette aventure. Ta présence à mes côtés a été essentielle, et je t'en suis profondément reconnaissante.

Et à vous, cher lecteur, que vous soyez venu chercher un simple extrait ou que vous ayez eu la patience de me lire jusqu'au bout, je vous remercie sincèrement pour l'attention que vous portez à mon travail.

Table des matières

Liste des abréviations

C2 : Command and Control, ensemble des fonctions de commandement et de coordination permettant à un commandant de planifier, diriger et contrôler les opérations et forces armées.

CCIR : Commander's Critical Information Requirements, ou informations critiques requises pour le commandeur.

DIMA : Détecter, Informer, Mémoriser, Agir, modèle de traitement de l'information.

DISARM : Disinformation Analysis and Response Matrix, matrice de lutte contre la manipulation de l'information, basée sur tactiques, techniques et procédures.

ELM : Elaboration Likelihood Model, théorie de la persuasion qui explique comment les attitudes sont formées et changées.

FOMO : Fear Of Missing Out, ou peur de manquer quelque chose d'important.

FRR : Force de Réflexion Rapide.

IA : Intelligence Artificielle.

LMI : Lutte contre les Manipulations de l'Information.

MBS : Military Brain Science.

MICE : Money, Ideology, Compromise/Coercion, Ego/Excitement, modèle de recrutement d'agents ou d'espions.

MIST : Misinformation Susceptibility Test, test développé par l'Université de Cambridge pour mesurer la sensibilité à la désinformation.

NAFO : North Atlantic Fella Organization, communauté luttant contre la désinformation par l'humour.

NBIC : Nanotechnologies, Biotechnologies, technologies de l'Information et sciences Cognitives.

NeuroS/T : Neurosciences et Technologies, domaine interdisciplinaire.

NLP : Natural Language Processing, interprétation automatique du langage humain.

OODA : boucle OODA ou OODA loop : Observe, Orient, Decide, Act.

PTSD : Post-Traumatic Stress Disorder, ou syndrome de stress post-traumatique.

SA : Situational Awareness, ou conscience de la situation.

SAD : Système d'Aide à la Décision.

SMOG : Simple Measure of Gobbledygook, score de lisibilité utilisé pour estimer la complexité d'un texte à partir de la longueur des phrases et du nombre de mots à trois syllabes ou plus.

TDCS : Transcranial Direct-Current Stimulation, technique de stimulation cérébrale portable qui délivre un faible courant électrique au niveau du cuir chevelu.

UnCODE : Unplug, Corrupt, disOrganize, Diagnose, Enhance, modèle de classification des objectifs de la guerre cognitive.

UPFD : User Preference-aware Fake News Detection, framework de détection de la désinformation sur les réseaux sociaux.

VIGINUM : agence française de VIGIlance et de protection contre les ingérences NUMériques étrangères.

VR : Réalité Virtuelle.

Liste des tableaux

Liste des figures

Introduction générale

Dans notre monde ultra-connecté, de nombreux acteurs sont aujourd'hui en mesure d'exploiter les réseaux sociaux ou divers outils numériques pour influencer à grande échelle des individus ou des populations ciblées, parfois situés à l'autre bout du globe. Ces manipulations peuvent concerner l'inhibition d'actions ou au contraire le passage à l'action, le changement voire la radicalisation d'opinions, l'abrutissement d'une société et notamment de ses enfants, pour lui faire perdre ses capacités d'innovation et donc son avantage économique.

Ce phénomène, au cœur de notre étude, est désigné par le terme « guerre cognitive », ou cognitive warfare. Ce terme désigne un mode de confrontation souvent invisible et non déclaré, sous le seuil de la guerre ouverte, qui vise à manipuler et influencer les mécanismes cognitifs et notamment la prise de décision d'un adversaire. Gagliano (2016) suggère que « ce n'est plus celui qui possède la plus grosse bombe qui aura gain de cause dans les conflits du futur mais celui qui racontera la meilleure histoire ». Si la guerre ouverte existe toujours aujourd'hui, ce commentaire témoigne toutefois d'un intérêt croissant pour les récits et l'influence de l'opinion populaire.

La guerre cognitive est un concept émergent dans la littérature, en lien avec le perfectionnement des nouvelles technologies et le développement des connaissances sur la cognition, ainsi que l'implication de l'opinion publique dans les conflits. Employée pour déstabiliser l'adversaire, elle représente une menace stratégique de plus en plus préoccupante, qu'il convient d'apprendre à détecter afin de s'en prémunir ; d'autant plus qu'elle recourt à des techniques susceptibles d'affecter toute personne en interaction avec le monde numérique.

Ainsi, nous voyons émerger la nécessité de concevoir des systèmes d'aide à la décision capables de détecter, caractériser et contrer les actions de guerre cognitive qui visent les décideurs civils et militaires : ceux-ci constituent l'objet de cette thèse.

Nous étudions particulièrement les vulnérabilités spécifiques des décideurs civils ou militaires face à ce type d'attaque. En effet, en tant que figures d'autorité, leurs décisions, leur posture et leur leadership constituent une influence considérable. L'objet de cette thèse est d'établir les fondements préliminaires à la conception d'un système d'aide à la détection de ces stratégies et à la mise en place de contre-mesures.

Dans une première partie (Figure 1), nous proposons une définition de la guerre cognitive, ainsi qu'un axe de lecture de ses différentes dimensions. Nous évoquons également les enjeux posés par ce nouveau type de conflit. Enfin, nous décrivons des exemples de stratégies de guerre cognitive et des moyens utilisés (*Chapitre 1*). Nous abordons également la prise de décision et ses vulnérabilités qui peuvent être exploitées (*Chapitre 2*).

Dans une deuxième partie (Figure 1), nos travaux empiriques portent sur les facteurs accordant le plus de potentiel d'influence à un message écrit (*Chapitre 3*). Nous étudions également un exemple d'outil qui peut être utilisé pour l'aide à la détection et à la mise en place de stratégies de guerre cognitive, en prenant pour exemple un jeu de stratégie semi-collaboratif (*Chapitre 4*).

La troisième partie (Figure 1) propose une réflexion prospective sur un modèle d'identification des cibles potentielles de guerre cognitive, en prenant en compte à la fois l'influence potentielle de l'individu et ses vulnérabilités (*Chapitre 5*).

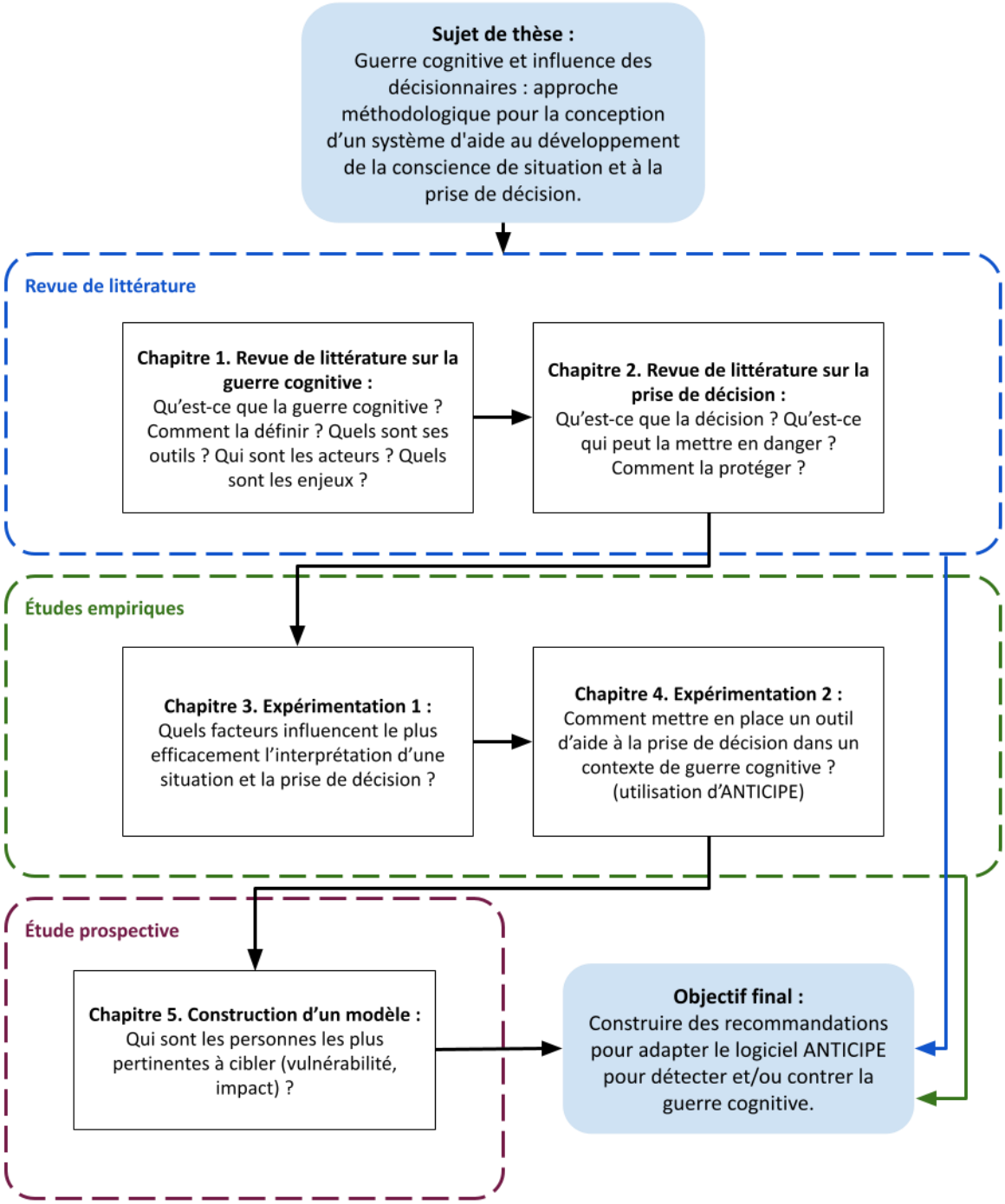

Figure 1 : Démarche et structure de la thèse

PARTIE I – Revue de littérature

Chapitre 1 - La guerre cognitive

La guerre cognitive étant un champ émergent dont la nature furtive rend difficiles les observations empiriques, la littérature présentée dans ce chapitre repose avant tout sur des contributions théoriques, avec peu de validation formelle. Cependant, nous estimons qu'elles constituent un consensus suffisant pour définir et structurer les concepts étudiés.

1.1 Historique

La guerre cognitive tire ses fondements chez Sun Tzu (V^{e} siècle) et Clausewitz (1832), deux penseurs de la stratégie militaire séparés par les époques et les cultures. Sun Tzu, dans *L'Art de la guerre (孫子兵法)*, prône la victoire sans combat en manipulant l'information et en exploitant les faiblesses psychologiques de l'adversaire. Clausewitz, dans *De la guerre* (*Vom Kriege*), décrit la guerre comme une extension de la politique visant à imposer sa volonté à l'ennemi et introduit les concepts de trinité clausewitzienne (la passion du peuple, le hasard militaire et la raison politique) et de « brouillard de la guerre » (Brittain-Hale, 2023). Cependant, ces auteurs n'avaient pas prédit l'utilisation de l'ingénierie sociale à grande échelle favorisée par les technologies numériques modernes, et ne mentionnent pas l'utilisation des technologies pour altérer la cognition humaine.

La Seconde Guerre mondiale a industrialisé la propagande de masse (radio, cinéma, affiches, tracts), afin de modeler l'opinion, soutenir la mobilisation et éroder la volonté de l'ennemi (Sostaric, 2019). Ces méthodes ont été perfectionnées pendant la Guerre Froide, qui voit émerger la guerre cognitive moderne (Bernal et al., 2020 ; Brittain-Hale, 2023) : une nouvelle guerre ouverte entre superpuissances promettait d'être extrêmement destructrice en raison de l'usage potentiel de l'arme atomique. Celles-ci ont donc mis en place des conflits détournés. Un exemple en est la guerre par procuration, ou proxy war, où des superpuissances soutiennent et opposent les uns aux autres des petits pays ou groupes armés (Mohlin, 2014). Un autre exemple est les actions de guerre informationnelle et de propagande, menées discrètement par les agences de renseignement des pays concernés.

Le pouvoir des récits et de l'opinion publique et leur importance dans un conflit sont connus depuis longtemps (Becker, 1978 ; Anderson, 1983). D'après Harbulot (2004), il existe une pression de la part de certaines populations occidentales pour éviter les conflits armés suite aux traumatismes des guerres du XX^{e} siècle ; certains affrontements sont alors transférés dans le domaine cognitif.

Ainsi, depuis les années 2000, nous observons une augmentation des actions de déstabilisation portée par le développement des technologies de l'information modernes (Porter, 2006 ; Prier, 2020). En effet, les états *« faibles »* rivaux des États-Unis, incapables de rivaliser avec eux dans un affrontement militaire direct, utilisent alors d'autres moyens : attentats, actions de déstabilisation, guérilla, etc. (Arreguín-Toft, 2001 ; Bühlmann, 2009). Ces actions sont menées par exemple par la Russie : tentatives de manipulation d'élections, propagande, cyberattaques, notamment envers les pays baltes, la France ou encore les États-

Unis (Antoniuk, 2024 ; Grynszpan, 2024). Elles sont aussi utilisées par la Chine pour affaiblir ses ennemis en sapant leur volonté de combattre (Orinx & Struye de Swielande, 2021).

Si la guerre cognitive peut se faire en l'absence d'affrontement direct, elle peut aussi accompagner un conflit armé pour semer la confusion, diminuer la résistance ou façonner l'opinion publique (Buchler, 2021). Par exemple, la Russie a mené contre l'Ukraine des opérations de guerre cognitive combinant cyberattaques, désinformation et piratage médiatique, comme en 2022 avec un deepfake du président ukrainien Zelensky incitant à la reddition, diffusé via une chaîne piratée (Wakefield, 2022). De même, la guerre cognitive a été utilisée comme préambule à l'offensive en Ukraine, avec des cyberattaques massives visant les infrastructures critiques pour éroder la confiance de la population ukrainienne envers son gouvernement, tout en diffusant un récit sur l'incompétence du gouvernement et la victimisation des russophones pour justifier l'agression (Malin, 2022 ; Takagi, 2022).

1.2 Définitions

Plusieurs définitions de la guerre cognitive sont proposées dans la littérature, traduisant des approches différentes (centrage sur les décideurs ou sur des populations, focalisation sur les effets ou sur les moyens employés, etc.). Ces définitions abordent plusieurs caractéristiques de la guerre cognitive, telles que la manipulation de la cognition humaine individuelle ou collective, l'exploitation des biais cognitifs, ou encore l'usage des technologies numériques à des fins d'influence. Nous en proposons ci-après un aperçu ainsi qu'une définition synthétique (paragraphe *1.2.3*).

1.2.1 Définitions de la guerre cognitive

Le terme de guerre cognitive est déjà utilisé depuis les années 2000 principalement en économie (Harbulot et al., 2002 ; Harbulot, 2004), mais la définition qu'en donne Harbulot se rapproche de la guerre informationnelle : elle est ainsi définie comme *« la capacité à utiliser les connaissances dans un but conflictuel »*, et l'École de Guerre Économique française mentionne l'obtention, la production ou l'entrave de certaines connaissances (Gagliano, 2016).

Depuis 2020, plusieurs définitions de la guerre cognitive, telle qu'elle est envisagée aujourd'hui, ont été proposées par différents auteurs. Notamment, en 2021, une conférence s'est tenue à l'École Nationale Supérieure de Cognitique (Bordeaux INP) en collaboration avec l'OTAN, portant sur le sujet de la guerre cognitive. Nombre des intervenants proposent des définitions de la guerre cognitive (Claverie et al., 2021).

Ainsi, dans l'avant-propos, le Général Philippe Montocchio (2021), directeur adjoint du Collaboration Support Office (CSO) STO, estime que la guerre cognitive est :

> *[Une forme de] manipulation permettant d'influer sur le comportement d'un individu ou d'un groupe d'individus, avec le but d'en tirer un avantage tactique ou stratégique. … Le cerveau humain devient le théâtre d'opérations. L'objectif est d'agir non seulement sur ce que pensent les individus-cibles, mais aussi sur la façon dont ils pensent, et en fin de compte, dont ils agissent.*

Claverie et Prébot (2021) ajoutent que la guerre cognitive peut consister à conduire des individus clés, civils ou militaires (soldats, techniciens, ingénieurs, décideurs, politiques) à *« avoir une représentation erronée du monde »* en s'appuyant sur les technologies de l'information. La guerre cognitive peut

s'attaquer aussi bien à la cognition individuelle qu'à la cognition collective. Ils donnent des exemples de manipulations et attaques du cerveau humain qui peuvent être combinées : « *Manipulation sémantique, illusion provoquée, distorsion perceptive, saturation de l'attention, trouble des apprentissages, de la mémoire de travail ou des souvenirs à long terme* ».

Claverie et du Cluzel (2021) intègrent la notion de biais cognitif (voir définition dans le paragraphe *2.2.1 Les biais cognitifs*). Ces auteurs précisent que le but est *« d'altérer les processus cognitifs d'ennemis, exploiter des biais ou des automatismes mentaux, provoquer des distorsions des représentations, des altérations de décision ou des inhibitions de l'action »* en présentant la réalité de manière biaisée, par exemple en la manipulant numériquement.

Claverie et du Cluzel (2021) décrivent aussi certains des objectifs de la guerre cognitive : *« conquêtes territoriales, influence (élections, troubles de population), perturbation des services publics (administrations, hôpitaux, secours, assainissement, eau ou énergie) ou de transport, effraction d'information (divulgation involontaire, publication de mots de passe…) etc. »*.

Du Cluzel (2020) souligne que la guerre cognitive affecte la faculté de comprendre et de produire de la connaissance, ainsi que l'esprit critique des individus, notamment en provoquant une forme d'épuisement psychologique.

Ainsi, la guerre cognitive brouille les frontières entre la guerre et la paix, permettant de mener des actions *« sous le seuil »*. Ses perpétrateurs peuvent l'utiliser pour parvenir à leurs fins sans déclencher de conflit armé. Elle peut même être menée contre ses propres alliés voire sa propre population. Orinx et Struye de Swielande (2021) le décrivent en parlant de guerre continue et de *« campagnes cognitives entre les guerres »*.

Au final, l'une des définitions qui nous paraissent les plus complètes est celle de Claverie (2021) :

> *Le cognitive warfare est l'un des moyens que des spécialistes utilisent pour modifier, orienter, et altérer la pensée humaine à des fins de conquête, supériorité ou inféodation des individus, ensemble d'individus, groupes ou populations. Il s'appuie sur la connaissance que l'on peut avoir des processus cognitifs que mobilisent ces individus dans l'utilisation et la maîtrise de leur environnement, notamment technologique, en ayant justement recours aux technologies numériques. De manière générale, il s'agit de modifier la conscience qu'ont les individus de la réalité pour leur faire prendre des décisions erronées ou les empêcher de prendre des décisions nécessaires. Le cognitive warfare est donc une pratique d'atteinte de la cognition à des fins de supériorité militaire.*

L'auteur ajoute que *« le cognitive warfare constitue un ensemble tridimensionnel (information, numérique et décision) »* pour atteindre *« la cognition de ceux qui conduisent, font ou évitent la guerre »*.

1.2.2 Objectifs de la guerre cognitive

Les objectifs inhérents à la guerre cognitive sont multiples, et ils peuvent être rangés dans deux catégories intimement liées : déstabiliser des personnes ou des groupes et influencer des décisions ou actions (les décisions considérées seront définies dans le *Chapitre 2*).

1.2.2.1 Déstabiliser des personnes ou des groupes

L'un des objectifs principaux de la guerre cognitive est de fragmenter et polariser la société, affaiblissant ainsi la cohésion nationale jusqu'à la diviser et la désorganiser, souvent en exploitant des différences ethniques ou en favorisant l'émergence de factions (Deppe & Schaal, 2024 ; Marjanović & Smiljanić, 2025). En outre, elle cherche à fragiliser et censurer la liberté d'expression, limitant ainsi la capacité des individus à communiquer et à s'exprimer librement (Cocron & Aronhime, 2022).

Un autre aspect crucial de cette stratégie est d'ébranler la confiance dans les institutions et la science, souvent par leur politisation ou par la propagation de discours catastrophistes, tout en délégitimant les alliances internationales, favorisant ainsi le nationalisme et l'isolationnisme (Cocron & Aronhime, 2022 ; Marjanović & Smiljanić, 2025). L'objectif ultime est d'amener l'adversaire à se détruire de l'intérieur, le rendant incapable de résister ou de prévenir la réalisation des objectifs de l'agresseur, par manque de souveraineté ou en exploitant le désordre comme diversion (Bernal et al., 2020).

La guerre cognitive peut également viser à provoquer une baisse de la productivité, comme le suggère Robin (2022) à travers l'exemple de TikTok (voir le paragraphe *1.6.3.1*). Elle peut aussi perturber des activités économiques stratégiques et des infrastructures essentielles. Raman et al. (2020) démontrent ce risque avec des simulations dans lesquelles une campagne de désinformation modifierait les comportements de consommation d'énergie des habitants d'une ville, entraînant l'effondrement de son réseau électrique.

Selon Du Cluzel (2020), ces actions sont conçues pour affaiblir, interférer et déstabiliser les populations, institutions et États ciblés, influençant ainsi leurs choix et minant l'autonomie de leurs décisions et la souveraineté de leurs institutions. Milstein (2020) va plus loin en décrivant une confusion profonde dans la perception collective de l'adversaire. Enfin, Tsikhelashvili (2022) souligne l'intention de démotiver, décourager et déstabiliser les individus visés, sapant ainsi leur volonté de se battre.

1.2.2.2 Influencer des décisions ou actions

La déstabilisation liée à des techniques de manipulation peut mener à l'altération de décisions ou encore l'inhibition d'actions.

Par exemple, les attaques relevant de la guerre cognitive peuvent saper la volonté de se battre d'une population ou d'une armée. Or, cette volonté est très importante dans un combat, au point où *« briser la volonté de combattre de l'ennemi tout en maintenant sa propre volonté de combattre est la clé du succès »* (Rand Corporation, 2018).

D'autres effets recherchés peuvent inclure la modification de la manière dont les individus perçoivent le monde, interprètent les informations et donc prennent leurs décisions, la modification des croyances et des jugements, la modification des comportements (Boyer, 2024 ; Cocron & Aronhime, 2022), ou encore l'influence ou la délégitimation d'élections (Deppe & Schaal, 2024 ; Antoniuk, 2024).

1.2.3 Proposition d'une définition du concept de guerre cognitive adaptée à nos travaux

Afin de fixer un cadre conceptuel, nous proposons une définition qui servira de référence pour ces travaux de thèse.

La guerre cognitive est un concept émergent qui vise à affaiblir l'ennemi afin d'obtenir un avantage tactique ou stratégique (Montocchio, 2021), notamment dans les domaines militaire et économique. Elle englobe diverses opérations menées contre l'esprit humain, ciblant les mécanismes cognitifs d'analyse et de traitement de l'information, ainsi que la construction des représentations mentales et décisions qui en découlent. Elle peut être menée et subie au niveau individuel ou collectif, à différentes échelles et à distance, par ou contre les populations, soldats, experts, ingénieurs, techniciens, groupes ou minorités d'opinion, d'ethnie ou de religion, entreprises, décideurs, responsables politiques, économiques, religieux, académiques ou militaires, etc. (Bernal et al., 2020 ; Claverie, Prébot & du Cluzel, 2021).

La guerre cognitive s'appuie sur les outils NBIC (Nanotechnologies, Biotechnologies, technologies de l'Information et sciences Cognitives), tels que les outils numériques et les réseaux sociaux, les substances chimiques, les illusions, la saturation de l'attention (Claverie et al., 2021), ou encore les biais cognitifs exploitables des cibles (Pinard Legry, 2022) pour altérer la cognition et les représentations du monde des victimes (Charzat, 2024).

La guerre cognitive s'inscrit souvent dans le temps long. Lorsqu'elle cible une population, les effets recherchés peuvent s'étendre sur une ou plusieurs générations, façonnant progressivement les représentations collectives et les comportements. Même lorsqu'elle vise un individu-clé, elle peut avoir pour objectif d'altérer ses capacités décisionnelles sur une période de plusieurs mois, voire années. Cette dimension temporelle introduit une limite aux expérimentations menées dans le cadre de cette thèse, qui ne peuvent en observer que les manifestations à court terme.

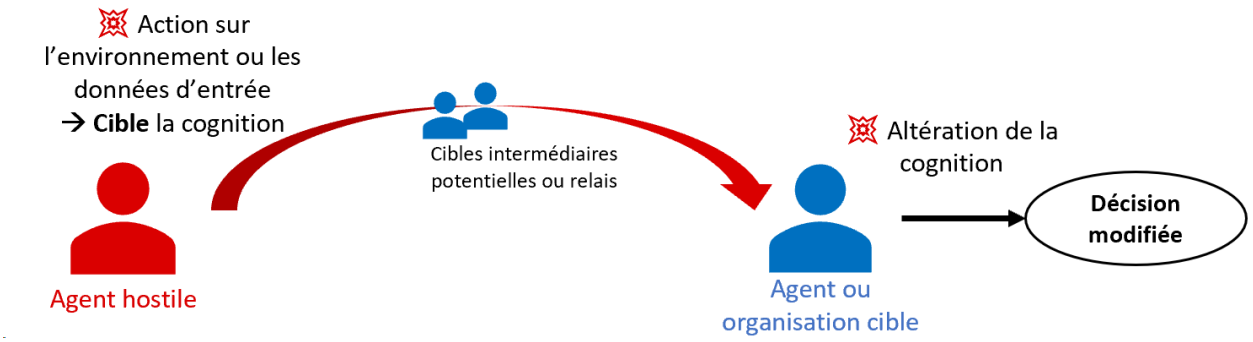

Figure 2 : Mécanisme général d'une action de guerre cognitive de l'agent hostile vers l'agent ou organisation cible

Les caractéristiques principales de la guerre cognitive sont les suivantes (Figure 2) :

- elle vise la cognition humaine ;
- elle a généralement pour objectif de déstabiliser, ébranler la confiance ou encore empêcher ou influencer des décisions (du Cluzel, 2020 ; Bernal et al., 2020) ;
- elle est souvent dissimulée, constituée de différentes actions en parallèles, de réactions en chaîne dont il peut être difficile d'identifier l'origine et l'objectif, dans le temps comme dans l'espace (Le Guyader & Cole, 2020) ;

- elle s'appuie généralement sur les outils technologiques et numériques modernes, qui facilitent l'accès aux cibles et permettent une diffusion rapide et à large échelle de campagnes de déstabilisation et d'influence.

> ➢ **Point-clé : définition de la Guerre Cognitive**
>
> Guerre cognitive : Stratégie dissimulée composée d'un ensemble d'actions facilitées par les technologies, entreprises par un acteur hostile pour influencer ou entraver les processus cognitifs d'un individu ou d'un groupe cible, dans le but d'orienter ses perceptions et ses décisions à son avantage.

1.2.4 Distinction par rapport à d'autres termes proches

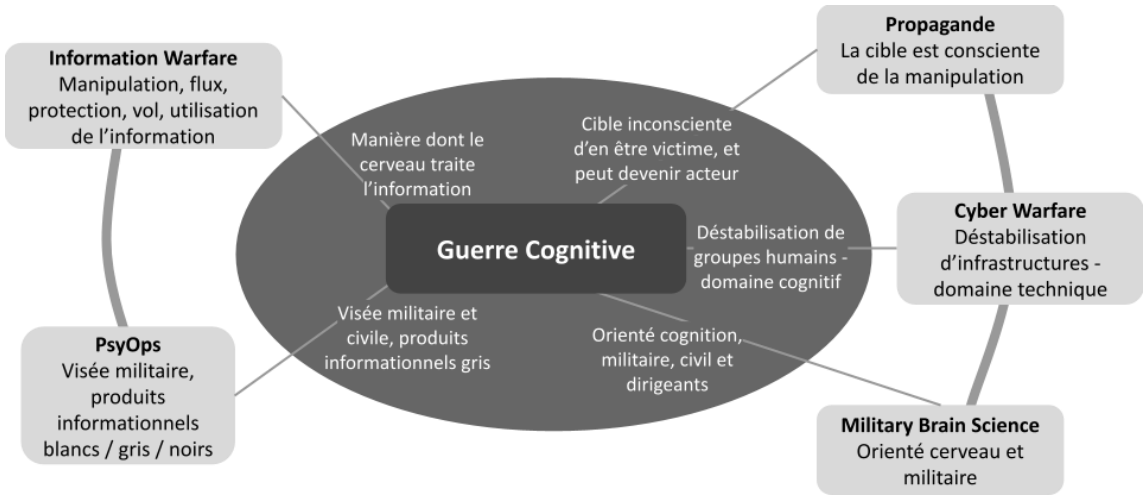

Figure 3 : Représentation schématique du positionnement de la guerre cognitive au regard des concepts proches

D'autres notions proches de la guerre cognitive (voir Figure 3) sont souvent évoquées dans la littérature, telles que l'Information Warfare, les opérations psychologiques (PsyOps) ou encore la propagande. Ces concepts, parfois utilisés de manière interchangeable, méritent d'être distingués. Nous proposons donc d'organiser et clarifier ces termes connexes.

La guerre cognitive est généralement un concept plus englobant, elle intègre plusieurs de ces pratiques et les transcende comme outils au service d'un objectif plus large (Montocchio, 2021).

1.2.4.1 Information Warfare, ou Information Operations

L'OTAN (NATO, s. d.) décrit la guerre informationnelle comme *« une opération conduite dans le but de gagner un avantage informationnel sur l'opposant »*. La guerre de l'information est centrée sur l'information, sa manipulation, ses flux, la manière dont on la protège ou la vole, et dont on s'en sert. Stuart Green (2008), ancien officier de la marine américaine, décrit l'information warfare comme composée de cinq éléments : la guerre électronique, les opérations sur les réseaux informatiques, les opérations psychologiques, les stratégies de tromperie militaire (déception) et la sécurité opérationnelle.

Comme on l'a vu dans le paragraphe *1.2 Définitions*, la guerre cognitive va plus loin : elle influence directement la cognition humaine et la manière de traiter l'information, de construire ou d'utiliser des connaissances (voir Tableau 1). Le renseignement étant un processus de construction de connaissances (Bulinge, 2010), il peut être affecté à la fois par la guerre informationnelle, qui influence les informations recueillies, et par la guerre cognitive, qui altère l'interprétation de ces informations.

Ravailhe (2024) estime que la guerre cognitive est connexe à la guerre de l'information, ce qui *« rendrait la guerre cognitive plus efficace au milieu de fake-news »*. En effet, d'après l'auteur elle repose sur la diffusion de faits authentiques soigneusement sélectionnés, conçus pour susciter des réactions et induire des biais cognitifs.

Tableau 1 : Différenciation entre la guerre de l'information et la guerre cognitive

Guerre de l'information	Guerre cognitive
Manipulation, flux, protection, vol, utilisation de l'information.	Manière dont le cerveau humain traite l'information et la transforme en connaissance, se représente le monde à partir de l'information à laquelle il est soumis, et ce que fait l'information au cerveau humain.

1.2.4.2 PsyOps, ou Psychological Operations

Les opérations psychologiques mettent en œuvre des produits informationnels qui sont des produits blancs (informations identifiables comme officiellement produits par la source qui les transmet), des produits gris (informations dont la source est ambiguë), ou des produits noirs (informations créées pour sembler provenir d'une source différente) (Jowett & O 'Donnel, 1986). La guerre cognitive utilise souvent des produits informationnels gris, dont la source paraît incertaine.

De plus, les PsyOps ont souvent une visée militaire, alors que la guerre cognitive peut s'attaquer à des populations civiles à large échelle, et *« tend à viser les infrastructures sociales civiles et les gouvernements »* (Whiteaker & Konen, 2021). D'après Trabelsi (2023), la guerre cognitive se sert d'outils tels que les manipulations développées pour les PsyOps.

Claverie et du Cluzel (2021) donnent quelques exemples d'opérations qui peuvent être menées dans le cadre de la guerre cognitive ou des PsyOps (*cf.* Tableau 2).

Tableau 2 : Différenciation entre les PsyOps et la guerre cognitive, d'après Claverie et du Cluzel (2021)

PsyOps	Guerre cognitive
Action sur les croyances, les perceptions faussées, l'illusion culturelle, les angoisses et les peurs, les faiblesses ou forces de personnalité, le refoulement… Pour influencer dans un certain objectif.	Action sur les cognitions, les dépassements sensoriels / perceptifs, la saturation attentionnelle, la tunnelisation attentionnelle, les erreurs de jugement, les biais cognitifs… Dans un but d'incapacité cognitive.

1.2.4.3 Propagande

La propagande est la *« transmission de communications, informations et messages… dans le but de produire des changements dans la conscience ou le subconscient de la population-cible, afin de changer les attitudes et comportements »* (Bobric, 2021). Elle est généralement menée ouvertement.

Tableau 3 : Différenciation entre la propagande et la guerre cognitive

Propagande	Guerre cognitive
Produits informationnels blancs : source de l'information clairement identifiable.	Produits informationnels blancs, gris ou noirs (l'information vient soit de la source effectivement déclarée, soit d'une source peu claire, soit d'une source totalement opposée à celle qui est déclarée).
Intentions généralement claires : la source cherche à se mettre en valeur dans ses stratégies de communication.	Intentions dissimulées et difficiles à identifier, recherche d'effets en chaîne et de rediffusion par des maillons non conscients de jouer un rôle dans cette chaîne.

La guerre cognitive est plus furtive que la propagande. En effet, pour cette dernière, il est facile de détecter l'origine de la campagne informationelle (elle utilise des *« produits informationnels blancs »* - Jowett & O 'Donnel, 1986) et la volonté de cette source de se mettre en valeur par une stratégie de communication (voir Tableau 3).

La guerre cognitive diffère également de la propagande dans le sens où tout le monde peut participer à la propagation d'informations ou de fausses informations diffusées dans un but de guerre cognitive, et donc devenir une arme sans en avoir conscience. La propagande pourrait donc également constituer un outil de la guerre cognitive dans certains contextes (Muhammad, 2024).

1.2.4.4 Cyber Warfare

Le cyber warfare, ou cyberguerre, désigne les opérations offensives et défensives menées dans l'espace numérique par des acteurs étatiques (avec une participation possible d'acteurs non-étatiques), visant à perturber, espionner ou endommager les systèmes informatiques, réseaux ou infrastructures critiques d'un adversaire. Elle peut impliquer des attaques par déni de service, l'injection de logiciels malveillants ou la manipulation de données (Robinson et al., 2015 ; Cornish et al., 2010).

Dans nos sociétés digitalisées et connectées, avec notamment l'internet des objets (IoT), de nombreuses fonctions critiques et infrastructures de contrôle dépendent du numérique d'une manière ou d'une autre : gouvernement, armée, institutions financières, fournisseurs d'énergie… Ainsi, les cyberattaques peuvent causer de nombreux dommages : perte de données, de temps, endommagement de matériel critique et jusqu'à des pertes humaines (Lehto, 2022 ; Bernal et al., 2020). L'exemple de Stuxnet est parlant : ce logiciel malveillant attribué à une opération conjointe des États-Unis et d'Israël permettait d'altérer la vitesse des centrifugeuses utilisées dans le programme nucléaire iranien, entraînant des défaillances inexpliquées et un retard considérable dans le programme (Fildes, 2010).

La guerre cognitive se rapproche de la cyberguerre, mais cette dernière est plus ciblée sur la maîtrise, l'altération ou la destruction des moyens informatiques ou des données. La cyberguerre relève donc plus

du domaine technique (Shaji et al, 2019), et l'effet cognitif est une conséquence, alors que pour la guerre cognitive, il est le but de l'action (Claverie & du Cluzel, 2021) (voir Tableau 4).

Tableau 4 : Différenciation entre la guerre cyber et la guerre cognitive

Guerre cyber	Guerre cognitive
Domaine technique. Déstabilisation d'infrastructures vitales ou stratégiques, principalement par l'utilisation de cyber-attaques.	Domaine humain. Déstabilisation de populations ou de personnes-clés par l'utilisation de différents types d'information et des particularités du cerveau humain qui les traite.

1.2.4.5 Military Brain Science

D'après Jin et al. (2018), *« la Military Brain Science (MBS) est une science innovante de pointe… basée sur les théories et les technologies de la médecine…, la biologie, la physique, l'informatique, les sciences militaires et de nombreuses autres disciplines »*. Elle a pour but de surveiller, protéger, lutter contre, réparer, améliorer le cerveau (Brunyé et al., 2020).

Jin et al. (2018) décrivent une division de 9 catégories d'actions militaires en lien avec le cerveau : *« comprendre le cerveau, protéger le cerveau, surveiller le cerveau, blesser le cerveau, interférer avec le cerveau, réparer le cerveau, améliorer le cerveau, simuler le cerveau et armer le cerveau »*.

Ils affirment également que cette science *« utilise l'application militaire potentielle comme boussole »* : ce sont donc des applications et outils qui peuvent être utilisés pour la guerre cognitive, mais restent centrés sur le domaine militaire (voir Tableau 5).

Tableau 5 : Différenciation entre la Military Brain Science et la guerre cognitive

Military Brain Science	Guerre cognitive
Science et applications tournées vers le cerveau du militaire en tant qu'organe central du système nerveux : sa surveillance, sa protection, son attaque, sa réparation et son amélioration.	Plus orientée vers la cognition et la pensée humaine que vers le cerveau physique. Peut concerner n'importe quel individu : militaires, civils comme dirigeants ou autres personnes clés, mais aussi des groupes (paragraphe *1.3.1 Cibles*).

1.3 Cibles & acteurs

Une des caractéristiques de la guerre cognitive est qu'elle peut être menée par de petits groupes d'acteurs voire des individus isolés, contre des cibles situées partout dans le monde. Elle peut être perpétrée, sous différentes formes, contre des individus isolés aussi bien que sur des groupes ou de larges populations. Il est important de se pencher sur ces cibles et acteurs potentiels en vue de la création d'un système de support à la détection des actions de guerre cognitive et à la mise en place de contremesures. Cette étude sera approfondie dans le *Chapitre 5*.

1.3.1 Cibles

Nous proposons d'envisager trois types de cibles (Figure 4) : les grands groupes de personnes partageant une caractéristique commune : nationalité, opinion, ethnie, religion, etc. ; les petits groupes de personnes partageant un objectif commun : entreprises, équipes, forces armées, etc. ; les individus critiques : décideurs, hommes politiques, leaders, communicants, vendeurs ou influenceurs, experts, chefs militaires ou des différents groupes mentionnés ci-dessus, soit toute personne qui a une influence sur un groupe via ses décisions ou le regard du groupe sur cette personne.

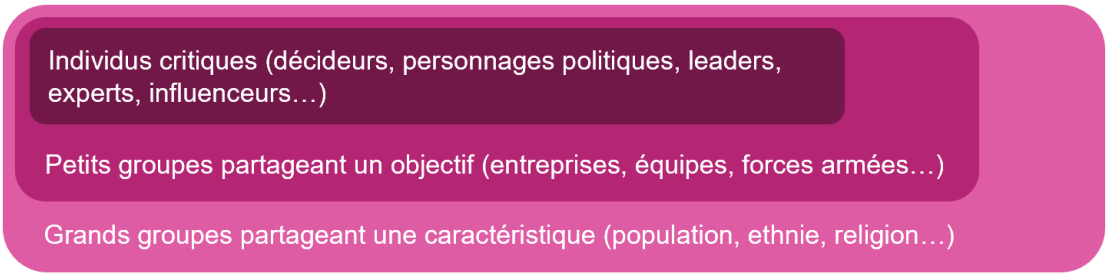

Figure 4 : Les trois types de cibles de la guerre cognitive.

Les personnes qui ont une influence (individus critiques) sur une décision ou sur un groupe de personnes, peuvent être visées en particulier. L'attaque peut être directe ou indirecte en visant le groupe auquel elles appartiennent. Cela peut avoir des répercussions sur ces individus critiques : par exemple, un changement d'opinion publique peut conduire à un changement de politique ou à une démission.

Comme le souligne du Cluzel (2020), *« n'importe quel utilisateur des technologies de l'information modernes est une cible potentielle. [La guerre cognitive] cible la totalité du capital humain d'une nation »*. Chacun peut y contribuer à divers degrés, consciemment ou non. Il ne faut pas négliger les tiers, individus neutres ou indifférents au conflit mais qui peuvent choisir un camp (Géré, 2005).

1.3.2 Acteurs

Comme les cibles, les acteurs de la guerre cognitive peuvent être variés. De nombreux outils de la guerre cognitive sont disponibles au grand public : information et désinformation, réseaux sociaux etc. (voir le paragraphe *1.6 Outils & armes de la guerre cognitive*). Même des individus isolés ou de petits groupes peuvent en tirer un pouvoir d'influence important avec peu de moyens et représenter une menace majeure pour les démocraties ou les opérations militaires (Hristakieva et al., 2022 ; du Cluzel, 2020).

➢ Bilan des paragraphes 1.1 à 1.3 : les points-clés

La guerre cognitive vise à affaiblir et déstabiliser l'ennemi en ciblant, souvent sur le temps long, sa cognition et sa prise de décision, au niveau individuel comme collectif. Il s'agit d'un concept émergeant notamment dans les domaines militaire et économique, facilité par les technologies de l'information modernes.

La guerre cognitive utilise certains outils des PsyOps et de la guerre informationnelle, mais elle s'en distingue par son ciblage de la cognition et de la manière dont l'humain traite l'information.

Toute personne connectée au monde numérique est une cible potentielle. On différencie notamment les grands groupes sociaux ou identitaires, les groupes organisationnels partageant un objectif commun, et les individus critiques, parmi lesquels les décideurs, à qui on s'intéresse en particulier dans ces travaux.

1.4 Représentations de la guerre cognitive

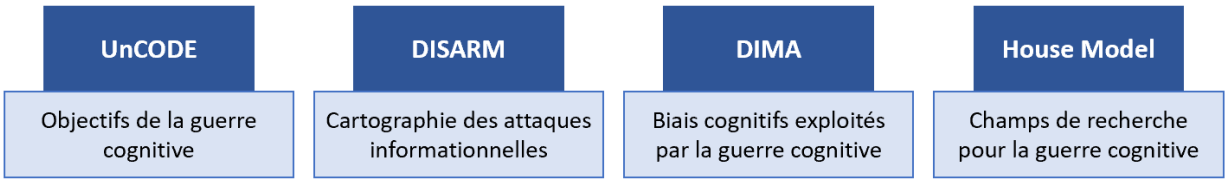

Figure 5 : Sélection de représentations de la guerre cognitive

Diverses modélisations et représentations ont été proposées pour décrire la guerre cognitive ou des actions proches (Figure 5). Nous en donnons ci-dessous quelques exemples qui aident à mieux la comprendre et pourront inspirer des modèles de détection à intégrer au système que cette thèse cherche à initier. Notamment, le système UnCODE classifie les objectifs de la guerre cognitive. DISARM est un outil construit pour la Lutte contre les Manipulations de l'Information (LMI), pour cartographier les attaques informationnelles. DIMA étudie les biais cognitifs exploités par la guerre cognitive. Le House Model s'intéresse en particulier aux champs de recherche pour la guerre cognitive.

1.4.1 UnCODE

UnCODE a été conçu comme une liste pour classifier les objectifs et les méthodes de la guerre cognitive (Ask et al., 2023) :

- Unplug (déconnecter) : neutraliser des individus, machines ou entités perçus comme des menaces cognitives (ex. : leaders d'opinion, analystes, IA). L'idée est d'éliminer les cerveaux d'une opération pour l'empêcher d'exercer son influence.
- Corrupt (Corrompre) : plutôt que viser les stratèges militaires, cette approche cible les exécutants qui mettent en œuvre les décisions. Exemples : provoquer une fuite des cerveaux, endommager les capacités cognitives (via drogues, stress chronique, etc.) d'une population.
- disOrganize (Désorganiser) : perturber les liens entre entrées et sorties cognitives. Par exemple, injecter de la désinformation pour fausser le raisonnement d'un stratège ou utiliser des agents biologiques pour modifier les comportements d'un groupe (impulsivité, prise de risques…).
- Diagnose (Diagnostiquer) : cherche à étudier les systèmes cognitifs cibles pour mieux concevoir les futures attaques : analyser les comportements sur les réseaux sociaux, les effets du stress, les types de contenus qui se propagent le plus rapidement, etc.
- Enhance (Améliorer) : augmenter les performances cognitives des alliés : neurostimulation, nootropes, augmentation sensorielle, recrutement de talents issus de la cible (ex. : fuite des cerveaux servant aussi à améliorer les capacités de la nation d'accueil).

Ces objectifs peuvent s'enchaîner : une action peut servir plusieurs buts UnCODE selon le contexte ou l'échelle temporelle. Par exemple, une technologie addictive peut à court terme désorganiser, et à long terme corrompre.

1.4.2 DISARM

Nous pouvons citer la matrice DISARM (www.disarm.foundation) pour la lutte contre la manipulation de l'information (LMI). Cet outil états-uniens, traduit en français par VIGINUM[1], est présenté comme capable d'établir un état des lieux des capacités, un diagnostic des forces et faiblesses, et de contribuer à bâtir une doctrine (Erard & Paquelet, 2024). La matrice est structurée par des tactiques, techniques et procédures qui composent une attaque informationnelle ou sur un système d'information. Celles-ci sont regroupées en phases (VIGINUM, 2024a) :

- Planification : planifier la stratégie et les objectifs, analyser les publics cibles.
- Préparation : élaborer les récits, fabriquer du contenu, établir la confiance…
- Exécution : diffuser les contenus, maximiser l'exposition…
- Évaluation : mesurer l'efficacité de la campagne.

La matrice DISARM est mise à jour régulièrement en fonction des retours de la communauté, afin de faciliter le partage de connaissances et la compréhension des techniques de manipulation utilisées par les différents acteurs, et ainsi améliorer les réponses.

1.4.3 DIMA

Inspiré des matrices DISARM et de MITRE ATT&CK (https://attack.mitre.org), une matrice utilisée en cybersécurité, le framework DIMA (voir Figure 6) a été conçu spécifiquement pour la guerre cognitive, afin de permettre *« d'identifier les tentatives d'utilisation de biais cognitifs dans notre consommation d'information »*. Il est composé de quatre étapes *« qui correspondent chacune à une phase du traitement de l'information reçue par une cible »* : Détecter, Informer, Mémoriser, Agir (Boyer, 2024).

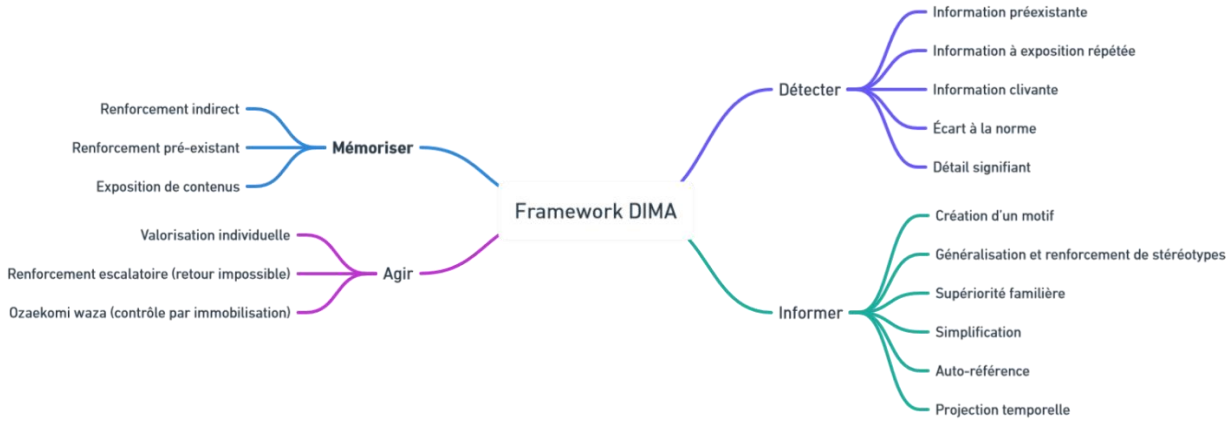

Figure 6 : Extrait du framework DIMA (voir le framework complet sur framindmap.org/c/maps/1457115/public*)*

- Détecter : cette séquence rassemble les biais cognitifs exploités pour rendre un message visible afin qu'il touche sa cible (effet de contraste, biais de distinction, biais de confirmation…).

[1] VIGINUM : agence française de VIGIlance et de protection contre les ingérences NUMériques étrangères.

- Informer : cette séquence rassemble les biais cognitifs *« utilisés pour orienter notre traitement de l'information et lui donner du sens »* (biais de confirmation, de représentativité, d'ancrage…).
- Mémoriser : cette séquence rassemble les biais cognitifs qui font que le message sera mieux mémorisé par la cible et intégré dans ses schémas de pensée et de représentation (cumul des biais des deux étapes précédentes pour mieux ancrer l'information).
- Agir : cette séquence rassemble les biais cognitifs exploités pour appeler la cible à l'action (biais de disponibilité qui peut pousser à *« surestimer l'urgence d'agir »*, biais de renforcement…).

1.4.4 House Model

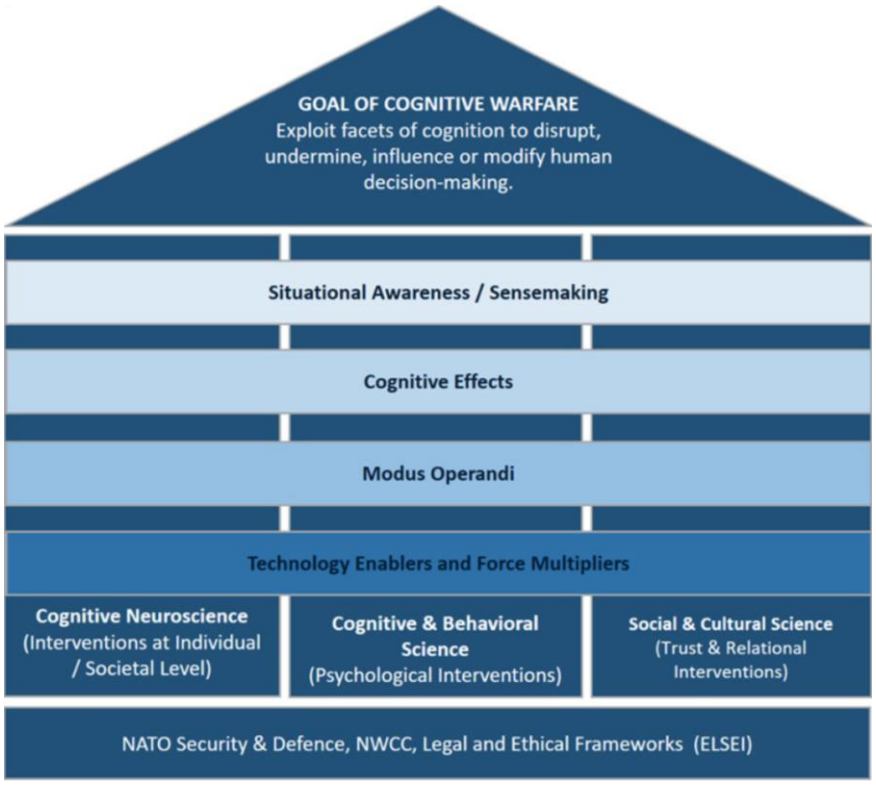

Figure 7 : House Model, d'après HFM Exploratory Team 356 (2023)

Le House Model (Figure 7) a été élaboré par la HFM Exploratory Team 356 (2023) de l'Organisation pour la Science et la Technologie (STO) de l'OTAN. Il sert de base pour l'élaboration d'une feuille de route stratégique en science et technologie afin de guider l'OTAN et ses partenaires dans leurs activités de recherche et leurs investissements en matière d'atténuation et de défense contre la guerre cognitive.

Ce modèle s'appuie sur trois piliers représentant les domaines prioritaires qui nécessitent des efforts de recherche : les neurosciences cognitives, les sciences cognitives et comportementales et les sciences sociales et culturelles. Le modèle se construit ensuite avec des éléments transversaux à ces piliers, qui permettent de contribuer à comprendre les attaques de guerre cognitive : la conscience de situation et le dégagement de sens, les effets cognitifs, les modus operandi et les possibilités technologiques.

1.5 Champs d'action de la guerre cognitive

Comprendre les champs d'action de la guerre cognitive nous permettra de mieux identifier qui pourrait utiliser un système d'aide à l'évaluation de la situation et à la décision dans ce contexte. Il s'agit de la première étape exploratoire pour définir les besoins de futurs utilisateurs.

1.5.1 Militaire

Comme son nom l'indique, la guerre cognitive est avant tout une forme de guerre, ce qui en fait un sujet d'intérêt pour le domaine militaire. Celui-ci a été pionnier en la matière avec les opérations psychologiques (PsyOps) et la guerre informationnelle. Ces opérations peuvent être considérées comme des composantes de la guerre cognitive mais elles n'en couvrent pas tous les aspects, notamment le ciblage direct de la cognition humaine.

1.5.2 Politique

La politique peut être une cible de la guerre cognitive, comme nous allons le voir avec les risques d'ingérences étrangères (ou par des acteurs internes) pour influencer des élections (voir le paragraphe *1.6.1.3 Manipulation d'élections*). Mais la concurrence entre différents partis politiques peut également voir l'apparition de stratégies s'apparentant à de la guerre cognitive : diffusion d'informations compromettantes ou de fausses informations à propos de concurrents (par exemple, les rumeurs selon lesquelles Barack Obama n'était pas né aux États-Unis – Wunder, 2021), stratégies de communication faisant appel à des biais cognitifs, etc.

1.5.3 Économie

Les acteurs économiques et industriels s'intéressent de près aux stratégies de guerre cognitive, comme nous l'apprend Nasi (2023) : l'École de Guerre Économique (EGE) française forme ses étudiants à l'espionnage industriel, au lobbying, à la désinformation, aux affrontements internationaux, etc. Ces thématiques répondent à une réelle demande du côté de l'industrie.

Dans ce domaine, l'enjeu de la guerre cognitive n'est pas nouveau, puisque dans les années 2000, Harbulot utilisait déjà ce terme, aussi bien pour décrire les affrontements entre pays qui défendent leurs intérêts qu'entre industriels concurrents et consommateurs au sein d'un même pays (Harbulot et al., 2002). Les outils de la guerre cognitive s'appliquent ici à affaiblir des concurrents, influencer les marchés, modifier la perception des consommateurs, influencer des décideurs politiques et économiques, asseoir le soft power des entreprises, etc. (de Mascarel, 2025 ; Gagliano, 2016).

Au niveau international, Gagliano (2016) décrit les marchés comme des instruments de pouvoir et donne l'exemple de *« l'utilisation de l'énergie comme arme de négociation [par] la Russie »*. Selon l'auteur, la guerre économique remplace d'autres formes de conflit, les gouvernements visant à développer un potentiel technologique, industriel et commercial pour renforcer leur économie et l'emploi local.

Il existe également des affrontements entre industriels et consommateurs. Les consommateurs peuvent exiger des entreprises une certaine responsabilité sociale et environnementale, les poussant à être attentives à leur image sous peine de perdre en compétitivité (Micheletti, 2003). Mais les industriels

peuvent aussi mener des stratégies de guerre cognitive envers le public pour conserver leur avantage, par exemple en contredisant les critiques scientifiques sur les dangers de leurs produits ou procédés industriels sur l'environnement ou la santé humaine (Oreskes & Conway, 2010).

Les pressions que les consommateurs peuvent exercer sur les industriels découlent parfois de stratégies de guerre cognitive mises en œuvre par certaines entreprises ou groupes d'intérêt pour nuire à l'image publique de leurs concurrents. L'image et la réputation d'une marque constituent en effet un capital stratégique crucial, capable d'influencer les performances commerciales et financières des entreprises (Fombrun, 1996). Or, les communications d'influence, souvent intégrées à des stratégies de communication ou de lobbying, sont légales et difficiles à identifier et à combattre. Elles mobilisent des registres discursifs éthiques ou moraux, ou encore des alertes fondées sur des risques environnementaux ou sociaux, parfois avec instrumentalisation d'ONG ou de groupes activistes qui favorisent une surenchère médiatique (Heath & Palenchar, 2009 ; Yaziji & Doh, 2009 ; Gagliano, 2016). Ces actions visent à affaiblir les entreprises, réduire leurs soutiens financiers et nuire à leur réputation, pouvant entraîner une chute significative de leur valeur boursière en peu de temps (Baumard, 2002).

Les politiques économiques peuvent aussi être façonnées par des campagnes d'influence visant à modeler l'opinion publique. Une réforme n'est acceptée que si elle repose sur une idéologie perçue comme légitime. Ce phénomène peut être exploité à des fins de guerre cognitive pour favoriser certains choix économiques au détriment d'autres (Rohrlich, 1987 ; Campos & Giovannoni, 2006).

1.5.4 Monde de l'entreprise

Dans le monde de l'entreprise, il existe de nombreuses manipulations cognitives, notamment par le biais du marketing, qui bien souvent manipule la clientèle potentielle (Danciu, 2014 ; Calo, 2013).

Les entreprises peuvent influencer les dynamiques politiques, notamment grâce à la collecte massive de données. Cambridge Analytica, un cabinet de conseil politique clandestin, a donné son nom à un scandale en 2016, dans lequel il est question d'interférence présumée dans l'élection présidentielle américaine de cette même année (Gayard, 2018). Cambridge Analytica a exploité les données d'une application Facebook, recueillies sous couvert d'un test de personnalité, pour obtenir des informations politiques sur environ 50 millions de personnes. Travaillant avec des personnalités politiques républicaines des États-Unis depuis 2012, l'entreprise est également suspectée d'avoir des liens avec WikiLeaks et des agents russes (Berghel, 2018). Ce scandale a démontré la possibilité de comprendre intimement les pensées des individus à travers leurs données (Prichard, 2021).

Cette implication d'entreprises dans les jeux d'influence n'est pas un cas isolé. Par exemple, l'officine israélienne Team Jorge a créé *« plusieurs milliers de comptes sur les réseaux sociaux »* pour diffuser des campagnes de désinformation ciblant des hommes d'affaires, des personnalités politiques, des lanceurs d'alerte ou encore des criminels présumés. Team Jorge aurait interféré dans plus d'une trentaine d'élections, principalement en Afrique, mais aussi en Europe, Asie du Sud-Est et Amérique latine. L'officine propose aussi des services de piratage de messageries d'adversaires (Leloup & Reynaud, 2023).

1.6 Outils & armes de la guerre cognitive

Pour construire un système d'aide à la décision capable de détecter les actions de guerre cognitive et proposer des contre-mesures, il faut identifier les moyens utilisés dans ces stratégies. Il est important d'en comprendre la nature, les usages et les effets. Cette analyse constitue un élément central du développement du système envisagé.

Tableau 6 : Classification d'armes et outils de guerre cognitive

Outils offensifs	Cibles privilégiées			Outils défensifs
	Civils	Leaders civils	Militaires	
Soft power, influencer l'opinion publique	X			Éducation, prévention, influence
Semer la panique, déstabiliser la confiance envers les gouvernements	X			
Manipulation d'élections	X	X		
Sanctions économiques	X	X		
Influencer l'opinion de leaders	X	X		*« Force de réflexion rapide »*
Cyberattaques	X	X	X	Lutte informatique défensive
Fausses informations / désinformation Réseaux sociaux Intelligence Artificielle	X	X	X	Éduquer à l'esprit critique et à la vérification de sources, médiation des réseaux sociaux, vérificateurs d'information
Armes biologiques (microbes, toxines), chimiques, acoustiques, nanotechnologies…	X	X	X	Détection et moyens de protection et de défense adaptés
Outils MBS (Military Brain Science) Augmenter le soldat Amélioration / stimulation du cerveau (militaire) Armes neuroS/T pour optimiser les performances (BCI, produits neuro-psychopharmacologiques, stimulation du cerveau, appareils d'augmentation neuro-sensorielle…)			X	Outils MBS défensifs : utiliser les mêmes outils qu'en offensif (augmenter le soldat…)
Biais de la décision		X	X	Processus protectifs, entraînement (OODA Loop), outils de partage de conscience de situation pour + de résilience
Inhibition de l'action	X	X	X	
Impact de la VR sur les comportements			X	Utiliser la VR en entraînement

Pour atteindre ses objectifs, tout acteur impliqué dans un conflit peut recourir à divers types de guerre, qu'elle soit cinétique, cyber ou autre. Kodalle et Ormrod (2023) soulignent que les effets cinétiques doivent toujours être pris en compte : par exemple, une personne atteinte d'une balle dans le cerveau voit évidemment sa cognition perturbée. Selon eux, l'influence cognitive peut avoir un effet plus subtil, durable et utile à long terme.

Il existe de nombreux outils et armes permettant d'obtenir ces effets ou de s'en protéger, notamment relevant des NBIC (Nanotechnologies, Biotechnologies, Informatique et sciences Cognitives). Nous nous intéresserons principalement à ceux qui visent à déstabiliser un adversaire et influencer sa cognition.

Nous proposons de les catégoriser en fonction de leur caractère (offensif ou défensif), leur champ d'application (populations civiles, leaders ou décideurs, militaires) et leur catégorie (nanotechnologie, biotechnologie, informatique, cognitique, économique, politique, processus et méthodes, physique). Le Tableau 6 présente une proposition de classification d'outils de guerre cognitive, en tentant d'opposer lorsque c'est possible, un outil offensif avec les outils défensifs qui peuvent contribuer à le contrer. Des précisions sur chaque outil mentionné dans ce tableau seront apportées ci-après.

1.6.1 Influence par la culture, l'économie et la politique

Caractère : offensif

Cibles privilégiées : civils, leaders

Catégorie : économique, politique

1.6.1.1 Soft power

Dès 1990, Joseph Nye décrivait le changement de paradigme du pouvoir : à la fin de la guerre froide, il devient apparent que les ressources ne déterminent plus seules la capacité d'une nation à obtenir ce qu'elle veut des autres. La culture et l'idéologie constituent d'importantes sources de pouvoir et d'influence (Nye, 1990 ; Vuving, 2009). Ainsi, le soft power désigne la capacité d'un acteur à influencer les autres sans contrainte, en rendant ses valeurs, actions ou politiques attractives ; contrairement au hard power qui impose, le soft power amène les autres à vouloir ce que l'on souhaite, par l'admiration ou la légitimité perçue (Gallarotti, 2011). Le soft power est une forme de guerre culturelle et de propagande, déployée dans une stratégie de long terme pour améliorer les capacités d'influence et l'image internationale des nations qui l'utilisent (Gazeau-Secret, 2013).

L'un de ses modes opératoires est la fenêtre d'Overton (Russell, 2006) : il s'agit d'aborder dans les médias certains sujets et opinions discutables ou qui ne font pas l'unanimité depuis un angle de vue (ou une *« fenêtre »*) acceptable, puis élargir cette fenêtre en rendant progressivement ces idées plus acceptables par le grand public (Lopez et al., 2021 ; Russell, 2006). Il s'agit d'un outil puissant, car il permet de *« changer radicalement l'opinion des gens, sans qu'ils réalisent qu'ils ont été habilement manipulés »* (Segura, 2018). De Morgny (2024) mentionne aussi *« l'endiguement cognitif »*, qu'il décrit comme *« la négation d'un sujet, d'un concept, la volonté de faire disparaître une notion, un aspect de la réalité »*, soit

en omettant le sujet en question dans le discours, soit en niant ou dénigrant sa pertinence (ironie, accusations de complotisme).

Le soft power peut également s'exercer par un mode de vie attractif ou des outils numériques. Gallarotti (2011) montre que la liberté, la tolérance et une qualité de vie élevée incitent d'autres pays à les adopter, ce qui crée une influence sans contrainte. Santini et al. (2018) introduisent le *« software power »* en tant que soft power, impliquant l'utilisation de bots politiques simulant des individus réels en ligne pour influencer subtilement l'opinion publique.

1.6.1.2 Sanctions économiques

Certaines sanctions économiques peuvent être considérées comme des stratégies de guerre cognitive, en ce qu'elles permettent de parvenir à ses fins sans affrontement armé et par une manipulation de l'opinion publique. Par exemple, les États-Unis ont tenté de renverser le régime socialiste de Castro à Cuba et le gouvernement chilien d'Allende sans usage de la force, par des sanctions économiques, pour ne pas aller contre l'opinion publique américaine. Ces tentatives n'ont pas été très efficaces à Cuba par méconnaissance de la population : celle-ci était habituée à une certaine sobriété et à des politiques économiques restrictives (Zylberberg, 1975). Elles ont été plus efficaces au Chili, où la classe moyenne reposait sur la société de consommation et a constaté que son gouvernement n'était pas capable de maintenir son niveau de vie en déclin. Cela a conduit au renversement du gouvernement chilien par un régime capitaliste, sans que les États-Unis ne soient tenus pour responsables (Brenner, 1990 ; Qureshi, 2008 ; Fischer, 2009).

1.6.1.3 Manipulation d'élections

Les interférences lors d'élections, par des acteurs étrangers ou domestiques, ne sont pas rares : acteurs russes, iraniens, entreprises diverses collectant des données d'utilisateurs ou vendant des prestations d'influence, etc. (Delaunay, 2024 ; Albertini et al., 2023 ; Salamanos et al., 2019). Plusieurs exemples d'ingérences de la Russie dans des élections sont donnés dans le paragraphe *1.7.1 Russie*.

Ces interférences ciblent les citoyens votants, afin d'influencer leur choix et de tenter de mettre au pouvoir une personne qui serve au mieux les intérêts des perpétrateurs. Dans ce but, les réseaux sociaux sont un outil très exploité : des entreprises comme Facebook *« ont identifié des opérations d'ingérence électorale russes et iraniennes hautement sophistiquées »*. En 2019, Facebook a annoncé avoir retiré quatre réseaux de comptes d'origine iranienne et russe de sa plateforme (Rosen et al., 2019). Les faux utilisateurs de ces réseaux diffusent de la désinformation et de la propagande, notamment pour influencer les élections (Khaled et al., 2018).

1.6.1.4 Lutter contre la manipulation d'élections

Pour contrer les ingérences étrangères dans les processus électoraux, plusieurs mesures peuvent être adoptées afin de préserver l'intégrité et la légitimité des élections.

Nous pouvons mentionner l'adoption de lois visant à renforcer l'intégrité des élections (Stewart III, 2022). Ces lois pourraient améliorer la transparence institutionnelle et l'accessibilité de l'information. Des efforts de communication pourraient également être engagés pour renforcer la confiance du public dans le

gouvernement et le système électoral (Stewart III, 2022 ; Norris, 2014). La lutte contre la fraude électorale, les cyber-intrusions et la corruption apparaît comme un élément crucial pour maintenir cette confiance, tout comme la protection de la liberté d'expression (Higashijima et al., 2024 ; Rothstein, 2013).

Une autre piste est d'améliorer l'éducation civique et la littératie politique via des programmes éducatifs et des campagnes de sensibilisation (Galston, 2001). Réduire la polarisation et favoriser le dialogue entre les groupes sociaux pourrait renforcer la résilience démocratique (Brady & Kent, 2022).

Backes et Swab (2019) proposent des stratégies spécifiques pour contrer les ingérences russes dans les pays baltes : identifier et inclure les populations les plus vulnérables, notamment les populations russophones, et leur proposer des informations nationales en russe et des programmes d'intégration pour les jeunes. Déterminer les vulnérabilités des systèmes d'information nationaux pourrait également contribuer à sécuriser les processus électoraux.

Bien sûr, une part importante de la lutte contre la manipulation d'élections consiste à lutter contre la désinformation (voir le paragraphe *1.6.4.2 Lutter contre la désinformation*).

1.6.2 Cyberattaques

Caractère : offensif

Cibles privilégiées : civils, militaires, leaders

Catégorie : informatique

Les cyberattaques constituent un outil de guerre cognitive : elles peuvent faire tomber des infrastructures essentielles pour les populations et ainsi leur faire perdre confiance dans leur gouvernement, empêcher celui-ci de communiquer pour rétablir la vérité (Malin, 2022) ou encore voler des informations, exposer des cibles à certains messages, etc.

Le vol d'informations peut être critique à différents niveaux : obtenir un avantage économique ou stratégique, mais aussi faire chuter la confiance des usagers dont les données ont été volées envers l'organisme qui n'a pas su les protéger (banque, hôpital, institution publique…), notamment si l'attaque est accompagnée d'une campagne de communication ou de désinformation (Strzelecki & Rizun, 2022).

Les données exposées sur les réseaux sociaux, couplées à celles issues de fuites de données *« circulant sur des forums d'attaquants »*, permettent à des personnes mal intentionnées d'identifier les cibles potentielles et d'accéder à ces dernières, de *« déterminer les biais exploitables chez leurs futures victimes pour construire des scenarii cohérents »* et de rendre l'attaque plus crédible, la cible croyant que la personne est légitime du fait des informations qu'elle connaît (Jolicard & Gardin, 2024).

Sur le même thème, le Dr James Giordano mentionne, lors de sa conférence de 2018 pour le Modern War Institute, la possibilité de modifier les données médicales d'un individu, pour changer son traitement, la manière dont elle est perçue par les assurances, ou encore divulguer des informations personnelles sensibles. Cela permettrait de manipuler une personne, son usage, le rendre incapable ou invalide pour le service, etc. Il est possible de cibler ainsi des individus-clés ou même des groupes pour changer la manière dont ils sont traités et perçus, ce qui affectera leur santé et leur stabilité.

1.6.3 Réseaux sociaux

Caractère : offensif, défensif

Cibles privilégiées : civils, leaders, militaires

Catégorie : informatique, cognitique

1.6.3.1 Offensif

Les réseaux sociaux sont un outil de guerre cognitive particulièrement puissant, étant donné qu'ils permettent de diffuser des messages très rapidement, à de larges communautés, sans qu'il soit toujours facile d'en retracer la source. D'autant plus que 38% des français s'informent uniquement via les réseaux sociaux (Newman et al., 2024). L'influence peut être volontaire, par la diffusion de messages ciblés, mais aussi involontaire, inhérente au fonctionnement même des réseaux sociaux et de leurs algorithmes de recommandation (Margraff, 2024). Ceux-ci proposent à l'utilisateur des contenus ou des comptes à suivre calqués sur ses préférences, qui lui donnent la fausse impression d'être maître de ses choix (Janin, 2024).

Il est important de noter que les réseaux sociaux sont aussi une menace pour l'armée, qui n'est pas coupée du monde et compte dans ses rangs de nombreuses jeunes personnes « *issues de la catégorie la plus sensible aux réseaux sociaux* ». Outre les nombreuses influences potentielles, les individus concernés peuvent aussi voir leur capacité d'attention diminuer (Janin, 2024).

Baughman et Singer (2023) rapportent que le ministère de la Défense de la République Populaire de Chine (www.mod.gov.cn/gfbw/jmsd/4931739.html) identifie quatre tactiques de manipulation sur les réseaux sociaux visant à reconfigurer le cadre mental global à travers lequel une population perçoit et comprend les événements :

- La perturbation de l'information : diffusion d'informations ciblées via des comptes officiels sur les réseaux sociaux pour influencer la compréhension d'une situation (par exemple un conflit) par le public cible.
- La compétition discursive : utilisation de trolls et de récits biaisés s'appuyant sur les biais émotionnels et la connaissance du fonctionnement des algorithmes des réseaux sociaux pour propager une vision du monde spécifique, tout en rendant les utilisateurs imperméables aux informations contraires.
- La coupure de l'opinion publique : saturation des réseaux sociaux avec un seul récit dominant en utilisant des bots, l'automatisation et l'IA, pour créer de la confusion, nourrir l'anxiété et le doute ou encore ébranler la confiance. Il est possible d'utiliser le machine learning pour repérer les personnes les plus vulnérables émotionnellement et les cibler avec du contenu émotionnellement chargé.
- Le blocage de l'information : utilisation d'attaques cyber, de blocages d'accès à l'information voire destruction physique des infrastructures de communication de l'adversaire pour l'empêcher de diffuser ses propres informations.

Les réseaux sociaux permettent en effet de diffuser des rumeurs, de semer le doute grâce à de fausses informations semblant provenir de sources fiables, avec des faux comptes parfois très élaborés et réalistes (voir le paragraphe *1.6.4 Fausse information, désinformation & information*), ou avec de vrais comptes piratés. Si les bots augmentent souvent la portée des posts en créant des messages et repartages automatisés, ils ne sont pas nécessaires : une étude du MIT suggère que la surprise et le dégoût, plus fréquemment inspirés par la désinformation, contribuent à une large diffusion de celle-ci par les humains (Vosoughi et al., 2018).

Les réseaux sociaux sont souvent accusés de générer des *« bulles »* ou *« chambres d'écho »* (Spohr, 2017) : chaque utilisateur est confronté à des contenus qui lui plaisent et le confortent dans ses idées. Ce mécanisme permet de conserver l'utilisateur le plus longtemps possible sur la plateforme, grâce à une sensation de confort intellectuel, sans remettre en question ses idées. Mais *« le danger réside [alors] dans la possible fragmentation de la société en une myriade de petites bulles séparées les unes des autres et heureuses de l'être. … Chacune des bulles est susceptible d'être déstabilisée ou perturbée au moindre contact »* (Cao et al., 2021). De plus, lorsqu'il y a une confrontation avec l'opinion opposée, elle ne conduit généralement pas quelqu'un en désaccord à se remettre en question (Mercier, 2017).

Un autre danger des réseaux sociaux est qu'ils permettent d'occuper leurs utilisateurs avec des contenus futiles qui réduisent les capacités d'attention, de mémoire et de raisonnement des utilisateurs, amenuisant ainsi l'esprit critique de larges communautés (Le Guyader, 2020). Cela relève de l'exploitation de la dopamine et l'économie de l'attention, ou *« théorie du Brain drain »*, à savoir l'accaparement constant d'une partie des capacités d'attention d'un individu en raison d'une connexion permanente par peur de *« manquer quelque chose d'important »*, ou FOMO – Fear Of Missing Out (Ward et al., 2017). L'addiction à internet et aux réseaux sociaux pourrait même entraîner une atrophie des structures cérébrales impliquées dans le traitement cognitif, d'après une étude en neuroimagerie sur les effets de l'addiction à Internet sur le cerveau (Takeuchi et al., 2018). En témoigne le mot de l'année 2024 désigné par Oxford : *« Brain rot »*, soit le *« pourrissement de cerveau »* (Yazgan, 2025). Ces médias vont jusqu'à entrer en compétition avec le sommeil de leurs utilisateurs, comme l'affirmait le patron de Netflix lui-même en 2018 (Gayard, 2018).

Ainsi, Robin (2022) affirme que l'application TikTok utilise divers outils pour retenir l'attention de ses utilisateurs ; contrairement à son équivalent DouYin qui propose aux jeunes chinois de nombreux contenus éducationnels, TikTok endormirait l'esprit de ses utilisateurs les plus fragiles et détournerait les jeunes de l'éducation et l'apprentissage (notamment des sciences) durant une période cruciale de leur développement. L'auteur craint des *« conséquences sociales et économiques importantes »* : réduction du potentiel d'innovation, augmentation des divisions. Su et al., dans une étude de 2021 sur les fondements neurologiques de l'attrait de TikTok, concluent que les algorithmes de recommandation influencent l'activité cérébrale pour maintenir l'attention des utilisateurs sur le contenu suggéré, et l'utilisation de TikTok serait liée à une moindre maîtrise de soi et des comportements problématiques.

1.6.3.2 Défensif

Comme évoqué dans le paragraphe *1.6.3.1* avec l'exemple de DouYin, lorsqu'ils sont bien régulés, les réseaux sociaux peuvent contribuer à l'éducation et à l'intérêt des populations pour des domaines utiles à la société. Ils peuvent créer un espace pour un apprentissage informel des sciences et promouvoir leur étude, comme l'ont montré Battrawi et Muhtaseb (2013) dans une étude de cas sur la page Facebook

« Creative Minds ». De plus, *« Pékin a limité l'utilisation des réseaux sociaux et jeux-vidéo [pour] réallouer le temps et l'attention disponible des jeunes chinois vers l'apprentissage »* selon Robin (2022). En démocratie, cette régulation peut constituer un danger pour la liberté d'expression.

La rééducation ou la désintoxication numérique a été largement étudiée pour les personnes qui souhaitent réduire leur dépendance ; d'après Janin (2024), elle pourrait être une piste pour les militaires, accompagnée d'un entraînement pour rééduquer leurs capacités d'attention.

1.6.4 Fausse information, désinformation & information

Caractère : offensif, défensif

Cibles privilégiées : civils, leaders, militaires

Catégorie : informatique, cognitique

La désinformation est la diffusion intentionnelle de fausses informations pour tromper. Elle est souvent liée à la propagande politique ou aux stratégies militaires. Elle se distingue de la mésinformation, qui est une diffusion involontaire d'informations erronées ou mal interprétées, sans volonté de nuire (Grace & Liang, 2023).

La désinformation et les efforts pour la contrer attirent l'attention de publics variés. Cela ne l'empêche pas de se répandre à large échelle, car elle repose souvent sur des doutes préexistants, est plus ou moins crédible et parfois difficile à détecter. Ces caractéristiques en font une arme puissante de déstabilisation.

Il est important de noter que si une campagne de désinformation peut combiner des informations réelles, déformées, exagérées ou fabriquées de toutes pièces, certains auteurs soutiennent que la guerre cognitive peut se passer de fausses informations (Hübert & Little, 2020 ; Cao et al., 2021). Une communication reposant sur de vraies informations bien choisies, comme des informations compromettantes, permet *« d'alimenter une polémique pertinente vérifiée par des faits objectifs »*, rendant la conspiration plus difficile à démontrer (Gagliano, 2016). Ces informations peuvent notamment être issues *« d'intrusions ciblées pour collecter des informations sensibles et organiser des fuites »* (Bertrand, 2023a).

1.6.4.1 Offensif

Les campagnes de désinformation exploitent les vulnérabilités des publics cibles, en s'appuyant sur leurs anxiétés et croyances préexistantes (Lewandowsky et al., 2017). Une autre faiblesse exploitée par la désinformation est le cerveau humain en lui-même : il prend des raccourcis, ou heuristiques, qui ne sont pas efficaces pour déterminer la fiabilité des messages. De plus, il est possible de croire à des déclarations simplement parce qu'elles sont accompagnées de preuves, sans les vérifier : les fausses informations peuvent ainsi s'appuyer sur des images falsifiées, ou même des images authentiques mais provenant d'un contexte différent (Newman & Schwarz, 2024 ; Lazer et al., 2018).

La désinformation est plus efficace lorsqu'elle provient de sources internes au groupe ou jugées crédibles, évoque des émotions fortes (peur ou indignation par exemple), dévalorise un groupe opposé plutôt que

le groupe ciblé par la désinformation lui-même, et lorsqu'elle est répétée fréquemment, même si elle contredit ce que le groupe cible sait déjà (American Psychological Association, 2023).

La désinformation, portée par les réseaux sociaux, peut présenter des risques bien réels. Ils ont participé à la popularisation des dispositifs TDCS (Transcranial Direct-Current Stimulation) et des applications thérapeutiques sur smartphone, qui contribuent à prendre en charge des syndromes comme les troubles de l'attention, le stress post-traumatique (PTSD), la dépression ou l'anxiété. Si ces techniques offrent des perspectives médicales prometteuses, leur accessibilité croissante inquiète : certains TDCS sont reproduits artisanalement ou commercialisés sans régulation. Ce sont des outils et méthodes disponibles sur le marché, facilement accessibles, ciblant les jeunes et les populations à risques. Or, non approuvés par des spécialistes, ils peuvent avoir des conséquences néfastes : irritation, confusion, modification de la cognition et de l'humeur. Des puissances hostiles pourraient disséminer via des applications thérapeutiques ou les réseaux sociaux, de fausses instructions pour la construction de TDCS ou des conseils volontairement erronés, ciblant des individus en état altéré (Bernal et al., 2020 ; Day et al., 2022).

Cependant, les réseaux sociaux ne sont pas nécessaires à la propagation de la désinformation. Par exemple, dans un contexte de conflit armé, elle peut s'immiscer dans les communications entre les protagonistes. Un exemple a eu lieu dans le cadre d'une opération de guerre électronique menée par Israël pendant la guerre des Six Jours, en juin 1967. Après avoir brouillé les radars égyptiens, des opérateurs israéliens maîtrisant l'arabe égyptien ont infiltré le réseau de communication radio de leur défense aérienne. Ils ont transmis de faux ordres et annulé les ordres légitimes, semant la confusion et empêchant les égyptiens d'utiliser efficacement leurs communications radio. Ces actions ont contribué à la défaite des forces aériennes égyptiennes (Black & Morris, 1991).

Parmi les moyens offensifs utilisant de vraies informations, Hübert & Little (2020) étudient le Kompromat, un procédé qui consiste à obtenir des informations compromettantes sur une personne, soit en ciblant des personnes déjà compromises et en récoltant des preuves, soit en les poussant à se compromettre.

1.6.4.2 Lutter contre la désinformation

La lutte contre la désinformation est complexe car *« il est plus facile de tromper les gens que de les convaincre qu'ils ont été bernés »*, et il est difficile de résister à *« la facilité de consommer ou de retransmettre des messages ou des informations qui vont dans le sens de nos convictions »* (Pinard Legry, 2022 ; Riant, dans Bertrand et al., 2023). Cependant, des solutions et outils de lutte sont proposés, dont certains sont décrits dans le Tableau 7.

Tableau 7 : Avantages et inconvénients de divers moyens de lutte contre la désinformation

Moyens de lutte contre la désinformation	Avantages	Inconvénients
Supprimer les faux comptes et débunker les fausses informations & contre-information	Débunk : peut aussi contribuer à éduquer le grand public contre la désinformation et le rendre moins vulnérable à de futures attaques.	À petite échelle seulement, les débunks sont moins viraux que les fausses informations et touchent moins de monde, manque de confiance en les vérificateurs de faits
Efforts d'inclusion des groupes et communautés exclus	Efficace sur le moyen à long terme si c'est bien mené.	Coûteux et pas de garantie d'efficacité
Education à l'école (enseigner à vérifier la source de l'information)	Efficace	Prend du temps, nécessite des enseignants compétents
Réduire la sensibilité à la désinformation via des jeux éducatifs	Ludique, engageant, peut être efficace	Possible à petite échelle seulement, ou à inclure dans l'éducation scolaire
Média d'information sur la désinformation	Leur influence s'accroît sur les réseaux sociaux	Portée limitée, diffusion généralement moins efficace que celle des fausses informations
Sites d'identification et de référence de fausses informations	Aident à vérifier la source et la véracité d'une information, notamment pour les journalistes	Il faut un effort de recherche et de vérification de la part des lecteurs
Vérificateurs de faits	Sensibilisent le public sur les réseaux sociaux, peuvent aider les journalistes	
Outils d'analyse des messages sur les réseaux sociaux	Possibilité d'utilisation à large échelle	Besoin d'analystes humains pour la vérification et l'efficacité
Sanctions légales	Dissuasion, mesures préventives	Ressources nécessaires pour identifier les coupables, ils ne sont pas toujours attaquables en justice (pays différent…), interrogations éthiques (liberté d'expression)

Détecter la désinformation :

Comme indiqué dans le Tableau 7, il est possible de lutter contre les fausses informations en les traquant, que ce soit via la répression ou en publiant des contre-informations (Aro, 2016). Mais encore faut-il que la contre-information soit partagée aussi largement que l'infox, ce qui est rarement le cas (Vosoughi et al., 2018), et que le public visé perçoive la source de la contre-information comme étant fiable (Bateman & Jackson, 2024). Une piste pour répondre à ces difficultés est l'utilisation de l'humour, comme le fait la communauté NAFO (https://nafo-ofan.org) : *« leur approche n'est pas seulement ludique, elle est stratégique : occuper le terrain des réseaux sociaux »* (Cvitkovic, 2024). En effet, il est possible d'être influencé par une fausse nouvelle même en sachant qu'il s'agit d'une désinformation (Akinwumi & Oladimeji, 2025). Lutter contre la désinformation sur les réseaux sociaux en les *« débunkant »* et en les supprimant, ou encore en supprimant les faux comptes, ne peut fonctionner qu'à petite échelle (McIntyre,

2023). Or, la désinformation atteint des dimensions très importantes, il faut donc trouver de nouveaux moyens de la combattre, et si possible de la prévenir.

Des outils permettent de détecter les faux comptes ou les bots sur les réseaux sociaux, en analysant les paramètres des différentes caractéristiques du compte (nom d'utilisateur qui ne représente rien par exemple), le comportement (pas de pause entre les interactions permettant à l'utilisateur de dormir), le contenu des messages postés ou encore les graphes représentant le réseau d'utilisateurs et leurs interactions (Wunder, 2021 ; Cresci, 2020 ; voir *Annexe 1*). Dou et al. (2021) s'intéressent à l'historique et aux engagements sociaux des utilisateurs qui diffusent des infox. Ils proposent un framework appelé UPFD, qui utilise les préférences des utilisateurs et la modélisation de graphes pour détecter la désinformation.

Une autre approche proposée par Bertrand (2023b) serait de lever l'anonymat sur les réseaux sociaux et de rendre les algorithmes plus transparents. Elle affirme que cela pourrait réduire *« les biais comportementaux qui alimentent en retour les possibilités de manipulation »*. Cependant, *« des mesures perçues comme excessives pour lutter contre la désinformation pourraient susciter la méfiance et générer des effets inverses à ceux souhaités »*. Dans ce cas, nous pourrions craindre d'attenter à la vie privée ou à la liberté d'expression.

Identifier les populations vulnérables :

Comme évoqué dans le paragraphe *1.6.1.4 Lutter contre la manipulation d'élections*, une part de la lutte contre la désinformation passe par l'identification des communautés exclues ou en marge de la population majoritaire du pays, et des efforts d'inclusion de ces groupes. En effet, ils pourraient avoir une tendance plus marquée à refuser le récit national et être plus vulnérables à la désinformation (Backes & Swab, 2019).

Alerter les lecteurs :

Kirchner et Reuter (2020) ont montré que les utilisateurs accueillent favorablement les avertissements affichés sur les fausses informations, en particulier lorsqu'ils sont accompagnés d'explications.

Aiguiser l'esprit critique des lecteurs :

L'éducation à l'école est un moyen puissant de lutte contre la désinformation, mais elle prend du temps et nécessite une formation des enseignants (McDougall et al., 2018).

Grace et Liang (2023) s'intéressent aux jeux éducatifs en s'appuyant sur deux théories de la communication : la théorie de l'inoculation, qui propose de préparer les individus à résister aux fausses informations comme un *« vaccin »*, et la théorie de la transportation narrative, qui affirme qu'immerger l'individu dans une histoire permet de provoquer un engagement émotionnel et cognitif et modifie plus efficacement ses croyances et attitudes. Grace et Liang proposent des jeux tels que :

- Harmony Square : le joueur incarne un « Directeur de la désinformation » dont le but est de semer le chaos, ce qui permet de mieux comprendre les mécanismes de manipulation.
- Lamboozled : les joueurs doivent convaincre les habitants d'une ville fictive de moutons de ce qui est vrai ou faux en s'appuyant sur des preuves.
- Fake You! : le joueur crée des titres crédibles pour une image, puis choisit le titre le plus crédible parmi trois. Il peut aussi construire une histoire cohérente à travers plusieurs titres.

Il existe des médias d'information sur les fausses informations, et la popularité des vérificateurs de faits augmente sur les réseaux sociaux. Wunder (2021) relève l'existence de *« divers sites web où les fausses informations sont identifiées en référençant la provenance et la source »* (www.politifact.com ; www.factcheck.org ; www.newsguard.com), sites utilisés par les journalistes pour éviter de diffuser de fausses informations. Ces derniers peuvent aussi être aidés par des vérificateurs de faits, afin que les médias traditionnels soient des sources sûres d'informations. Ces pratiques existent déjà, mais pourraient être généralisées.

Au niveau individuel, Slater et Rouner (1996) proposent de lutter contre les *« effets de bulles »* et la désinformation en se confrontant à des idées inverses aux siennes, en prenant du recul (surtout en cas de réaction émotionnelle), en se mettant à la place des autres, en s'interrogeant et s'informant sur la désinformation et les techniques de persuasion et en étant actif dans la manière dont nous consommons l'information. Ils soulignent également l'importance de filtrer les informations : il n'est pas nécessaire de lire tous les articles que nous rencontrons, et il est parfois bénéfique pour la santé mentale de se déconnecter et d'être sélectif.

Il existe de nombreuses autres approches : ateliers pour apprendre à débattre de manière argumenée, à analyser ses propres biais et processus de raisonnement, à identifier les intentions des émetteurs de l'information, jeux de rôle, etc. qui ne peuvent pas être étudiés de manière exhaustive ici.

1.6.5 Intelligence Artificielle

Caractère : offensif, défensif

Cibles privilégiées : civils, leaders, militaires

Catégorie : informatique

Le développement rapide des Intelligences Artificielles génératives ces dernières années n'a probablement échappé à personne : génération automatique de textes, d'images, de vidéos de plus en plus crédibles et difficiles à distinguer des contenus réels ou créés par l'humain. Ces outils ne peuvent pas être négligés, de par leur puissance, leur accessibilité et leur large utilisation par toutes sortes de personnes pour des usages variés. De plus en plus de questions sont posées directement aux intelligences artificielles de génération de textes, plutôt qu'à des moteurs de recherche (Pham et al., 2024).

1.6.5.1 Offensif

Begou et al. (2023) ont montré que ChatGPT pouvait être utilisé par des cybercriminels pour faire de l'ingénierie sociale (hameçonnage), l'outil les aidant à rompre la barrière de la langue et de la culture.

La diffusion des fausses informations est de plus en plus facilitée par les outils d'intelligence artificielle. Parmi ces outils facilitateurs, nous pouvons mentionner les deepfakes, vidéos générées par intelligence artificielle qui peuvent par exemple montrer une personne récitant un discours qu'elle n'a jamais prononcé : leur danger est évident, étant donné qu'il est possible de faire dire n'importe quoi à n'importe quelle personnalité influente. Elles peuvent être rendues encore plus réalistes grâce à des technologies d'imitation du timbre de la voix et de l'accent d'une personne (Nechaev & Kosyakov, 2024). Le risque lié aux visages générés par intelligence artificielle est tout aussi réel : il permet de créer de nombreux faux

comptes sur les réseaux sociaux et d'humaniser les bots pour leur donner plus de crédibilité. L'écriture par intelligence artificielle aide à la diffusion de fausses informations, créant des articles, posts et commentaires sur les réseaux sociaux beaucoup plus rapidement qu'une équipe humaine. Comme le souligne O'Neil (2016), certains modèles algorithmiques, qualifiés « d'armes de destruction mathématique », produisent des effets cognitifs massifs et invisibles : par leur opacité, leur échelle et leur biais, ils orientent les décisions humaines, renforcent les inégalités et sapent le débat démocratique. Ainsi, une seule personne pourrait générer des milliers de commentaires et de posts sur les réseaux sociaux, orientés pour supporter ou discréditer une cause. Par exemple, le média social X (anciennement Twitter) accueille de nombreux bots, qui peuvent *« poursuivre des objectifs malveillants tels que l'ingérence dans les élections et la propagande extrême »* (Feng et al., 2021 – voir *Annexe 1*).

1.6.5.2 Défensif

L'intelligence artificielle est un outil à double tranchant : elle permet aussi la détection de contenus trompeurs ou falsifiés (deepfakes, bots, campagnes de manipulation) par analyse sémantique, détection d'anomalies ou encore avec la reconnaissance d'images (Bontridder & Poullet, 2021).

Elle peut également aider à surveiller les réseaux sociaux en temps réel et en continu, détectant des campagnes de désinformation et d'ingérence étrangère, notamment en période électorale, en analysant les tendances et les schémas de diffusion de l'information (Saade, 2025 ; McGovern, 2021). Elle contribue à modérer les contenus présentant de fausses informations (Alaphilippe et al., 2019).

L'IA peut être utilisée pour développer des outils éducatifs et des campagnes de sensibilisation visant à renforcer la résilience des individus face à la désinformation. En identifiant les biais cognitifs exploités par les campagnes de manipulation, elle contribue à élaborer des stratégies pour les contrer (Trilateral Research, 2025).

1.6.6 Armes NeuroS/T

Caractère : offensif, défensif

Cibles privilégiées : militaires, leaders, civils

Catégorie : biotechnologie, physique, cognitique, informatique

Les neurosciences et technologies (NeuroS/T) offrent de nouvelles possibilités opérationnelles pour la guerre, le renseignement et la sécurité nationale. Elles permettent *« d'augmenter les capacités des forces alliées »* ou *« affecter les capacités cognitives, émotives et physiques de l'adversaire »*. De plus, elles peuvent être appliquées *« dans des engagements cinétiques ou non cinétiques, pour produire des effets destructeurs ou perturbateurs »* (DeFranco et al., 2019 ; Kumar & Dixit, 2019).

DeFranco et al. (2019) décrivent diverses approches neurotechnologiques actuelles ou émergentes utilisées dans le cadre du renseignement et de la sécurité nationale. Celles-ci incluent l'amélioration des performances humaines grâce à des agents pharmacologiques (stimulants, euphoriques, modulateurs de l'humeur), à des dispositifs de neuromodulation (stimulation transcrânienne, interfaces homme-machine) et à des outils d'intelligence artificielle inspirés du fonctionnement cérébral. Ces technologies peuvent être utilisées pour affaiblir ou manipuler des adversaires, notamment à travers des substances

psychotropes (tranquillisants, hallucinogènes), des agents biologiques (virus, bactéries, édition génétique), des neurotoxines naturelles, ou encore des systèmes à énergie dirigée et nanomatériaux neuroactifs capables d'agir directement sur les fonctions cognitives.

1.6.6.1 Applications aux populations civiles et aux leaders

Les armes NeuroS/T peuvent avoir des applications aux populations civiles et à leurs leaders, dans des buts de déstabilisation ou de manipulation.

Krishnan (2017) introduit les notions de *« Drugs, Bugs, Bytes, Waves »*, des catégories qui décrivent différentes méthodes de manipulation cognitive, émotionnelle ou comportementale.

Drugs & Bugs - les armes chimiques, biologiques et neurotechnologiques :

La catégorie *drugs & bugs* englobe les calmants, hallucinogènes, incapacitants biologiques et chimiques, capables de ralentir la pensée, perturber le sommeil, altérer les émotions ou générer la peur. Si ces moyens se développent, nous pourrions voir leur diffusion massive, par exemple via des vecteurs biologiques comme les insectes ou les virus (Pinard Legry, 2022 ; Moreno, 2012).

Certaines armes chimiques et biologiques visent à influencer les émotions, pensées ou comportements de cibles humaines, comme l'illustre le Dr James Giordano dans une conférence donnée au Modern War Institute (2018). Il décrit une situation fictive où l'usage de microdoses de substances neuroactives (drogues ou toxines) permet d'altérer les pensées, émotions ou comportements d'un leader ciblé. Administrées discrètement (via sa boisson, un objet ou son environnement), ces substances peuvent incapaciter la cible ou modifier son attitude, influençant ainsi ses partisans. Ces derniers pourraient alors adhérer à un comportement plus favorable à l'attaquant ou perdre confiance en leur leader.

Concernant les armes biologiques, Giordano évoque des agents neuro-microbiens provoquant des symptômes neuro-psychiatriques aigus pour semer la panique, des agents microbiologiques génétiquement modifiés pour causer de nouvelles formes de morbidité ou de mortalité, ou encore des insectes porteurs de neurotoxines lâchés dans des zones à forte densité (villes, aéroports, stades), causant des troubles neurologiques, psychologiques, et une panique amplifiée par la désinformation en ligne. Enfin, le Dr James Giordano mentionne les nano-neuroparticules, capables de provoquer des hémorragies cérébrales à grande échelle, simulant une épidémie d'AVC.

Bytes & Waves - la manipulation cognitive par données et signaux :

Outre les agents chimiques et biologiques, Krishnan (2017) identifie une autre catégorie : *« Bytes & Waves »*, comprenant les données neuronales, les signaux subliminaux (images, sons) et autres technologies d'influence ou de thérapie, comme la stimulation magnétique transcrânienne (TMS) et la stimulation transcrânienne à courant continu (TDCS) qui peuvent être utilisés pour le traitement des symptômes de la dépression (Downar et al, 2024 ; Bennabi & Haffen, 2018). DeFranco et al. (2019) identifient deux principales menaces liées aux neuro-données : l'altération de la perception d'un individu ou groupe et la création d'effets ciblés de manipulation cognitive.

Nous pouvons donner l'exemple de l'ADS (Active Denial System), arme cognitive qui émet un rayonnement électromagnétique donnant une sensation de brûlure sous la surface de la peau (Buch & Mitchell, 2013). Par ailleurs, des recherches médicales expérimentent des implants électriques ou

optogénétiques, déjà testés chez l'animal, capables par exemple de créer de faux souvenirs de peur, et qui pourraient à terme être appliqués chez l'humain (Yu et al., 2020 ; Lau et al., 2020).

1.6.6.2 Applications militaires des armes NeuroS/T

Il existe de nombreuses applications militaires des NeuroS/T, notamment dans le cadre de la MBS, définie précédemment dans le paragraphe *1.2.4.5 Military Brain Science*.

Armes NeuroS/T offensives :

Jin et al. (2018) mentionnent plusieurs facteurs de risque pour le cerveau humain dans les contextes militaires : les champs magnétiques, les ondes de choc dues à des explosions, le stress prolongé lié à des opérations de navigation dans un espace confiné, ou encore les troubles psychologiques post-traumatiques, qui affectent l'état mental et l'efficacité au combat. Bien que ces effets soient initialement considérés comme secondaires ou indésirables, ils pourraient être délibérément recherchés. Les auteurs estiment que de nouvelles armes peuvent être conçues pour cibler le cerveau humain, exploitant des vecteurs variés comme le son, la lumière, les lasers, les explosions, les champs magnétiques etc. pour agir sur des zones cérébrales spécifiques.

Parmi les méthodes évoquées figurent les techniques dites *« smokeless »*, c'est-à-dire non létales, invisibles et silencieuses, permettant d'interférer directement avec le fonctionnement cérébral (Jin et al., 2018 ; Moreno, 2012) :

- *« Agents incapacitants du système nerveux »* ou du corps, y compris des armes perturbant les ondes cérébrales ;
- *« Armes à infrasons »* conçues pour *« interférer avec les tissus du cerveau et provoquer la folie par résonance »* ;
- *« Tactiques psychologiques »* destinées à *« interférer avec les croyances et pensées de l'ennemi, lui causant des blessures psychologiques et affaiblissant sa volonté de combattre »*
- Sons artificiels (bips, bruits spéciaux), capables de générer la peur, la dépression ou la confusion, réduisant ainsi les capacités opérationnelles.

Ces technologies s'inscrivent dans une *« guerre des cerveaux »* : un nouveau paradigme militaire où l'objectif est d'influencer la pensée, de perturber la prise de décision et de redéfinir le champ de bataille, en plaçant le cerveau humain au cœur des opérations offensives (McCreight, 2024). Elles sont des causes possibles du Syndrome de la Havane (voir le paragraphe *1.7.4 Syndrome de La Havane*).

Soldat & humain augmenté :

Les avancées en neurotechnologie et en biotechnologie cognitive offrent de nouvelles possibilités d'hybridité humain-système pour restaurer (par exemple traiter le syndrome de stress post-traumatique), renforcer (*« stimuler et améliorer les fonctions cognitives et physiologiques »*, améliorer la vitesse de traitement de l'information ou encore l'endurance), voire remplacer et dépasser les capacités humaines (adaptation à de nouveaux environnements par exemple), en particulier dans les environnements militaires (Johns Hopkins University & Imperial College London, 2021 ; Giordano, 2018).

Ces technologies s'inscrivent dans la logique des *« Bytes & Waves »* de Krishnan (2017). Selon Pinard Legry (2022), elles incluent les *« systèmes à énergie dirigée (micro-ondes, acoustique, électromagnétique) »* et les implants, qui pourraient *« établir une véritable interface entre la machine et le cerveau »*.

Les interfaces cerveau-ordinateur (BCI) sont déjà une réalité. Si leur image renvoie parfois à des dispositifs encombrants et invasifs, ce domaine connaît des avancées rapides. Un exemple marquant est l'implant Neuralink proposé par Elon Musk, conçu comme *« une intervention peu invasive qui pourrait moduler certains réseaux et fonctions du cerveau »* (Shaima et al., 2024). Ces dispositifs ne sont pas uniquement destinés au personnel militaire. Ils pourraient être appliqués à des civils influents ou des leaders d'opinion, avec des objectifs de performance ou de contrôle.

L'utilisation de différents outils des NBIC pour l'amélioration de la performance humaine est un sujet traité depuis longtemps (Roco & Bainbridge, 2003 ; Jin et al., 2018). Mais elle pose des questions éthiques, notamment relatives à la responsabilité de personnes qui seraient sous l'influence de la biotechnologie cognitive, au consentement pour l'utilisation de technologies invasives qui pourraient causer des dommages, ou encore à la vie privée (protection des pensées et souvenirs intimes des utilisateurs) (Johns Hopkins University & Imperial College London, 2021).

1.6.7 Cognitique et biais cognitifs

Caractère : offensif, défensif

Cibles privilégiées : civils, militaires, leaders

Catégorie : cognitique

La cognitique est une branche applicative des sciences cognitives qui s'intéresse au traitement automatique de la connaissance (Le Blanc, 2018). Elle peut être utilisée comme arme de guerre cognitive, notamment en exploitant les biais cognitifs des cibles.

Nous faisons ici une introduction aux armes cognitiques qui peuvent être utilisées pour la guerre cognitive (Tableau 8).

Tableau 8 : Armes cognitiques et biais cognitifs pour la guerre cognitive, d'après Claverie et du Cluzel (2021)

Outils offensifs	Effets	Outils défensifs
Promouvoir des outils numériques ou dispositifs qui atteignent les processus cognitifs / exploitation de failles numériques ou des interfaces des outils numériques d'aide ou de surveillance.	Possibilités de manipulation via ces outils.	Prévention, surveillance des pratiques populaires sur les réseaux sociaux.
Saturer la prise d'information / pollution attentionnelle, altérer la construction de représentations, induire des décisions inadéquates, paralyser la prise de décision.	Inhibition de l'action par indécision ou surcharge cognitive, tunnelisation attentionnelle. Mauvais ajustement de l'objectif attendu.	Jeux vidéo : amélioration de la concentration sur les détails, meilleur sens spatial, traitement multitâche efficace, meilleure capacité à travailler sous pression.
1e couche du cerveau (perception, automatismes…) → illusions visuelles, saturer l'attention, exploiter les automatismes.	Manipulation de la perception de l'environnement / de l'information, des actions.	
2e couche du cerveau (contexte, émotions, analyse, mémoire…) → implanter des souvenirs, déclencher des réflexes émotionnels par imposition de souvenirs, perturber la mémoire, exploiter les influences émotionnelles et les interférences.	Manipulation des émotions et du comportement.	Auto-contrôle ou contrôle partagé, analyse métacognitive d'anticipation (jeu et simulation), analyse des performances et retour d'expérience (« retex » dynamique)
3e couche du cerveau (conceptualisation, métacognition…) → biais de haut niveau (ambiguïtés de sens, manque ou excès de signification, ambiguïtés sémantiques…), empêcher la réalisation des raisonnements par pression temporelle, parasitage ou facilitation des erreurs de raisonnement	Manipulation de la réflexion, des choix, des actions…	
Exploitation des dissonances cognitives pour introduire des absences de cohérence entre modèles conceptuels et connaissances personnelles	Création de vulnérabilités cognitives ou de troubles psychopathologiques exploitables	
Biais d'auto-conviction	Fausses informations, fausses controverses, révisionnisme, contestation de la science, etc.	
Exploitation des tendances à l'induction et l'abduction (erreurs cognitives) pour les rendre plus fréquentes / probables	Erreurs de jugement, interprétation inadaptée de la situation et de ses causes	Mettre en place une vérification déductive, détecter les failles des raisonnements et procédures

1.6.7.1 Offensif

La cognitique peut accroître le brouillard de la guerre chez l'adversaire, altérer ses capacités cognitives en l'empêchant de dormir ou en lui causant par exemple des maux de tête. [Elle joue]

sur les peurs, l'hystérie collective des sociétés dans le but de faire exploser le contrat social. (Pinard Legry, 2022).

En témoigne l'exemple potentiel du *Syndrome de La Havane*, décrit dans le paragraphe *1.7.4*.

Outils numériques et technologiques :

La promotion d'outils numériques ou de dispositifs qui atteignent les processus cognitifs peut constituer une arme de guerre cognitive, comme nous l'avons vu avec l'exemple des TDCS et applications thérapeutiques dans le paragraphe *1.6.4.1 Offensif*.

Il existe plusieurs mécanismes permettant de perturber la cognition de l'ennemi. Ils incluent la saturation ou pollution informationnelle qui agit comme distraction, la désinformation, la surcharge cognitive (d'attention ou de décision)… Ils peuvent mener à la modification des représentations mentales, une interprétation inadaptée de la situation, des décisions et actions perturbées ou paralysées, voire des troubles de la personnalité. Par exemple, la tunnelisation attentionnelle est un phénomène dû à la surcharge cognitive où l'individu peut ignorer les alertes voire les désactiver, entraînant parfois des accidents (Bell et al., 2005 ; Claverie & du Cluzel, 2021 ; David & Bode-Asa, 2023). Ou encore, le syndrome de persévération désigne la mobilisation d'efforts vers un objectif unique, même si cet objectif est dangereux ou inadapté ; il se manifeste en particulier en cas de stress ou de charge de travail importante, et plus il dure, plus il est difficile d'en sortir (Dehais et al., 2003).

Connaissance du cerveau humain et neurologie :

Les sciences neurologiques offrent aussi des moyens de manipulation. D'après Claverie (2021), le cerveau est composé de trois couches qui ont chacune leurs spécificités et leurs vulnérabilités : le premier niveau, responsable des réflexes, est facile à leurrer avec des illusions visuelles par exemple ; le deuxième niveau, lié à la mémoire et l'affectivité, peut être manipulé en modifiant les souvenirs ou en provoquant des réactions émotionnelles ; le troisième niveau, qui traite du sens et de la métacognition, peut être perturbé par des ambiguïtés de sens, des conflits d'interprétation ou des informations contradictoires. Il existe des personnalités cognitives variées : certaines personnes privilégient les informations sensorielles plutôt que les aspects émotionnels ou mémoriels, ou se concentrent davantage sur les détails que sur l'ensemble. Connaitre cette diversité humaine permet de mieux comprendre sa cible et comment utiliser les incohérences entre ses connaissances et ses modèles mentaux, facilitant la perturbation de sa cognition, sa mémoire ou ses comportements (Walsh et al., 2006 ; Courbet & Benoit, 2013 ; Claverie, 2021).

Exploitation des biais et failles cognitives :

Les biais cognitifs ou failles cognitives sont des erreurs systématiques dans la manière dont nous percevons, jugeons ou prenons des décisions, résultant de raccourcis mentaux (voir le paragraphe *2.2.1 Les biais cognitifs*). Ils sont inhérents à la pensée humaine et constituent à la fois un avantage évolutif permettant de prendre des décisions plus rapidement, et une fragilité qui mène à des erreurs de jugement. Ils peuvent alors être exploités dans le cadre de la guerre cognitive. Par exemple, les informations peuvent être présentées de manière à déclencher ces biais via leur contexte, le format, les métaphores utilisées, les couleurs, les images, le langage (positif, négatif, émotionnel…), etc. (Tversky & Kahneman, 1981 ; Thibodeau & Boroditsky, 2011). Le stress, la pression temporelle, la surcharge informationnelle et cognitive, la fatigue, l'inconfort ou encore la présence de distractions dans

l'environnement peuvent augmenter l'usage d'heuristiques pour accélérer la décision, et donc entraîner plus d'erreurs cognitives (Hammond, 2000).

L'être humain est sujet à des erreurs de raisonnement, notamment l'abduction et l'induction, deux formes de raccourcis cognitifs pouvant générer des biais. L'abduction tire des hypothèses probables à partir d'observations, mais sans garantie de vérité, et sans vérification systématique. L'induction, au contraire, généralise à partir d'exemples, au risque d'introduire de fausses croyances ou d'ignorer des exceptions. Nous pouvons comparer ces mécanismes au concept d'attibution causale qui nous vient de la psychologie sociale : les individus cherchent spontanément à expliquer les comportements par des causes internes ou externes, mais peuvent alors tomber dans des biais tels que l'erreur fondamentale d'attribution, soit la tendance à surestimer les causes internes comme la personnalité, les intentions ou les attitudes (Heider, 1958 ; Ross, 1977 ; Weiner, 1985). Ces raccourcis mentaux, accentués par la rapidité de pensée et l'absence de vérification, peuvent être exploités pour manipuler ou induire en erreur (Fischer et al., 2014 ; Chemero, 2023 ; Claverie, 2021).

1.6.7.2 Défensif

Parmi les pistes pour améliorer la résistance aux biais cognitifs nous pouvons mentionner les jeux vidéo, qui permettent de développer de nouvelles compétences : amélioration de la mémoire de travail visuospatiale, de la vitesse psychomotrice, de la vitesse de traitement cognitif, de l'attention, de la concentration sur les détails, du sens spatial, de la capacité à travailler sous pression, etc. (Bediou et al., 2018 ; Zioga et al., 2024).

La formation peut contribuer à améliorer le raisonnement logique et la prise de conscience des biais et de leurs conséquences, mais elle ne suffit pas à les éliminer. En effet, l'appareil cognitif humain reste biologiquement stable, peu influencé par l'expérience ou l'apprentissage, rendant tout le monde vulnérable aux biais. Il est possible de développer des stratégies de contrôle et d'analyse métacognitive, mais celles-ci deviennent peu efficaces en situation de stress ou d'urgence, où les biais ancrés réapparaissent (Claverie, 2021 ; David & Bode-Asa, 2023).

Une autre piste pour se protéger des erreurs cognitives est la mise en place de processus structurés : vérification par les pairs, analyse, détection systématique des biais, retours d'expérience, etc. (Korteling et al., 2023 ; Fasolo et al., 2025).

1.6.8 Stratégies transversales

Caractère : offensif, défensif

Cibles privilégiées : civils, militaires, leaders

Catégorie : processus

1.6.8.1 Mise en place de processus

Kozloski (2018) souligne que les avancées en sciences de l'information et en sciences cognitives au cours des dernières décennies ont montré la manière dont les limites humaines affectent les performances des organisations. Il apparaît donc important d'identifier les processus qui peuvent être mis en place pour préparer les militaires comme les salariés d'entreprises à faire face aux menaces de la guerre cognitive. Ainsi, Buchler (2021) indique que le renforcement de la résilience face aux agressions cognitives repose sur la formation des équipes et l'instauration de routines de travail normées. L'auteur appelle ainsi à atteindre une *« maturité cognitive technologique »*, qui combine l'intégration et la collaboration humain-système, vers une adaptation *« agile, adaptative et intelligente »*.

Concernant la prise de décision en contexte incertain, Lagadec (2010) propose l'outil *« force de réflexion rapide »* (FRR). Il s'agit de cellules de réflexion constituées à l'écart des structures conventionnelles, autorisées à explorer des idées non conformes ou décalées. Ce mécanisme favorise une pensée créative, libérée des pressions hiérarchiques. Il a notamment démontré son efficacité durant la crise des missiles de Cuba en 1962, lorsque, à l'initiative de Kennedy, un petit groupe indépendant élabora une alternative diplomatique qui permit d'éviter l'affrontement armé. Ainsi, la constitution d'une FRR dès le début d'une crise permet de mener une analyse approfondie de ses caractéristiques, de ses protagonistes et de ses évolutions possibles, tout en proposant des options diversifiées au décideur.

1.6.8.2 Entraînement

L'entraînement peut devenir un levier stratégique contre la guerre cognitive, mais aussi un outil pour façonner des opérateurs plus résilients, plus réactifs, voire plus offensifs. En effet, les limitations cognitives naturelles (stress, fatigue, peur) sont amplifiées dans l'environnement militaire (Rabat et al., 2025). Une étude de Haushofer & Fehr (2014) démontre d'ailleurs que le stress, fréquemment rencontré par les officiers, détériore la prise de décision et la perception du risque. Selon Kozloski (2018), les organisations militaires traitent désormais les compétences cognitives comme leurs équivalents physiques, en cherchant à standardiser les techniques et à encadrer la performance mentale. Le concept de préparation cognitive, ou *cognitive readiness*, prend alors tout son sens : Morrison et Fletcher (2002) la définissent comme une *« préparation mentale »* traitant *« les compétences, connaissances, capacités, motivations et dispositions personnelles »*, indispensables pour agir dans des environnements complexes et imprévisibles. Elle vise à préserver des fonctions clés : conscience de la situation, mémoire, apprentissage, métacognition, automatisation des décisions, résolution de problèmes, flexibilité, créativité, leadership et régulation émotionnelle.

Parmi les outils d'entraînement, Hardy (2024) met en avant les wargames et serious games. Les connaissances et la préparation de l'individu peuvent être renforcées aussi bien en concevant ces jeux qu'en y jouant en interaction avec d'autres apprentis.

La réalité virtuelle (VR) constitue également un levier puissant. Dans une interview avec Ananthaswamy (2016), Metzinger souligne que les expériences immersives en VR ont une influence plus forte et plus durable sur le comportement que d'autres supports. L'expérimentation de Kammers et al. (2009), où le cerveau peut croire qu'une main en caoutchouc lui appartient, illustre un mécanisme transposable à la VR. Ainsi, Slater et al. (2010) ont montré qu'une perspective à la première personne d'un corps virtuel peut induire une illusion de transfert de propriété corporelle, confirmée par des mesures physiologiques.

Selon Metzinger, dans ces environnements, les individus adoptent souvent les comportements attendus de leur avatar : un avatar de grande taille favorise une posture plus agressive ; un avatar ressemblant à soi-même plus âgé incite à épargner pour la retraite. Il ajoute que les effets de la VR persistent après l'expérience et peuvent modifier durablement les attitudes. Il apparaît donc possible que si l'utilisation de la VR était plus démocratisée, la création et la promotion de certains jeux ou plateformes en réalité virtuelles pourrait devenir un outil puissant de guerre cognitive en contrôlant les caractéristiques du personnage incarné ainsi que les comportements que celui-ci adopte.

1.6.8.3 Stratégies randomisées

L'adoption de stratégies aléatoires, ou randomisées, permet de renforcer l'imprévisibilité dans les opérations militaires. En évitant les schémas d'action prévisibles, les forces armées peuvent compliquer la planification et la prise de décision de l'adversaire, le contraignant à disperser ses ressources et à réévaluer constamment ses hypothèses ; il peut même perdre du temps et des ressources pour analyser et prédire un raisonnement qui est en réalité aléatoire. Par exemple, si un véhicule suit systématiquement le même itinéraire entre deux points, il devient vulnérable à des embuscades ou des attaques ciblées. Introduire une variabilité dans les trajets rend ces mouvements moins prévisibles et donc plus sûrs (Priebe et al., 2021).

> ➢ **Bilan des paragraphes 1.4 à 1.6 : les points-clés**
>
> Plusieurs modèles existants permettent de comprendre les logiques de la guerre cognitive, notamment ses objectifs (UnCODE), les biais cognitifs exploités à chaque étape (DIMA) ou encore les champs de recherche qui interviennent (House Model).
>
> La guerre cognitive agit notamment dans les domaines militaire, politique, économique et de l'entreprise : elle peut être utilisée dans toute forme d'affrontement visant à obtenir un avantage stratégique.
>
> Elle mobilise un large éventail de moyens : désinformation, cyberattaques, réseaux sociaux, IA, armes neurotechnologiques, biais cognitifs… Ces outils peuvent être offensifs ou défensifs et s'appuient souvent sur les technologies NBIC pour manipuler la cognition humaine.

1.7 Exemples de stratégies de guerre cognitive

1.7.1 Russie

« La goutte d'eau creuse la pierre, non par force, mais en tombant souvent », proverbe du KGB rapporté par Chavalarias (2024). Cette logique s'inscrit dans la tradition russe de la Maskirovka, une doctrine de dissimulation et de manipulation qui s'étend à toutes les dimensions des opérations d'influence (Claverie, 2023). Même en l'absence de conflit armé, la Russie met en place diverses stratégies de guerre cognitive, ciblant les pays voisins (pays Baltes, Ukraine), l'Occident, mais aussi l'Afrique avec la présence du groupe Wagner et ses successeurs. Ces actions visent à obtenir un avantage stratégique sur ses adversaires,

parfois lié au contrôle de ressources naturelles comme *« les diamants en Centrafrique, l'or au Soudan ou le pétrole en Libye »* (Bertrand et al., 2023, pp. 16-17).

Ces opérations relèvent à la fois d'un récit interne et externe. En Russie, le gouvernement renforce ses fondations idéologiques autour de *« valeurs russes traditionnelles »* et d'un discours de restauration de grandeur impériale pour consolider l'adhésion populaire et opposer la société à l'Occident (Shtepa, 2021). À l'international, la Russie cherche à façonner l'opinion publique étrangère. Untersinger et al. (2024) rapportent qu'en 2022, des influenceurs européens ont été approchés pour diffuser de la propagande prorusse, mettant en avant la puissance militaire de Moscou et chercher à inspirer la peur d'un conflit mondial. En 2023, VIGINUM a identifié un réseau de 193 faux sites d'information diffusant des contenus manipulés au bénéfice du Kremlin, désigné sous le nom de *« Portal Kombat »* (VIGINUM, 2024b). Selon Chavalarias (2024), ces opérations s'accompagnent de techniques de désinformation comme les opérations sous faux drapeau, consistant à usurper l'identité d'un groupe pour le discréditer. RT et Sputnik, chaînes d'information soutenues par le gouvernement russe, participent également à cette stratégie en diffusant un récit favorable à la Russie en Europe (Limonier & Audinet, 2017).

Les campagnes de guerre cognitive russes incluent aussi des ingérences dans les élections de plusieurs pays : États-Unis, Royaume-Uni, Allemagne, Italie, France ou pays Baltes (Eady et al., 2023 ; Badawy et al., 2019). L'affaire DNC Leaks de 2016 illustre ce phénomène : des hackers russes ont divulgué des courriels internes du Parti démocrate américain, ce qui a contribué à fragiliser la candidature de Hillary Clinton et à polariser l'électorat (Mueller, 2019 ; Baezner & Robin, 2017 ; Inkster, 2016).

Cette influence est particulièrement marquée dans les pays Baltes, où les proximités géographiques, historiques et culturelles avec la Russie sont exploitées pour exacerber les tensions, notamment en jouant sur le sentiment de discrimination des minorités russophones (Backes & Swab, 2019 ; Cheskin, 2024). La stratégie russe ne crée pas nécessairement de nouveaux clivages, mais s'appuie sur des fractures existantes qu'elle amplifie par des moyens variés : désinformation, propagande, mémoire collective, réseaux sociaux, bots, deepfakes. L'objectif est d'influencer les comportements et les décisions en exploitant les fragilités cognitives et identitaires des publics ciblés (Marsili et al., 2021).

1.7.2 Chine

La Chine est un autre acteur important dans la guerre cognitive, même si ses actions sont souvent considérées comme plus discrètes que celles de la Russie.

1.7.2.1 Culture de guerre chinoise

La pensée stratégique chinoise s'ancre dans une tradition millénaire, dont Sun Tzu est l'emblème le plus connu. Dans L'Art de la guerre (V^{e} siècle av. J.-C.), il énonce l'essence même de la domination par la ruse : *« Soumettre l'ennemi par la force n'est pas le summum de l'art de la guerre, le summum de cet art est de soumettre l'ennemi sans verser une seule goutte de sang »*. La stratégie chinoise s'inscrit dans une logique de long terme, fondée sur la connaissance fine de l'adversaire (Hartley III & Jobson, 2020). Sun Tzu insiste : *« Connais ton ennemi et connais-toi toi-même »*. *Les 36 stratagèmes*, recueil de ruses militaires rédigé entre les XIV^{e} et XVII^{e} siècles, est un autre exemple de texte qui trouve une résonance culturelle dans les stratégies chinoises actuelles. Ainsi, Orinx et Struye de Swielande décrivent la doctrine chinoise des *« Trois guerres »*, particulièrement adaptée à la guerre cognitive moderne et plus flexible que les approches

occidentales : la guerre psychologique renforce la cohésion des alliés, neutralise les acteurs indécis, et érode la volonté et le moral de l'adversaire ; la guerre de l'opinion publique utilise les médias et réseaux sociaux pour modeler la perception du conflit, maintenir le soutien national, affaiblir le moral des adversaires et influencer les pays tiers ; la guerre juridique consiste à chercher à respecter le droit international, construire un récit légitime autour de ses propres violations si elles ont tout de même lieu et dénoncer les violations adverses.

1.7.2.2 Recherche et innovation

L'Armée Populaire de Libération définit les opérations dans le domaine cognitif comme visant à influencer directement les émotions, le jugement et les comportements en agissant sur la cognition. Elle considère ce domaine comme central pour la guerre future, l'innovation dans ce domaine est une nécessité stratégique (Charon & Jeangène Vilmer, 2021 ; Kumar, 2024 ; Fenstermacher et al., 2023). La Chine mobilise pour cela des technologies telles que le Big Data et l'IA pour collecter des informations sur les individus ciblés et leurs vulnérabilités psychologiques, analyser les opinions publiques et évaluer l'impact des opérations (Hung & Hung, 2022 ; Beauchamp-Mustafaga, 2023).

La Military Brain Science (MBS) s'inscrit dans ce cadre d'opérations en intégrant des recherches avancées en neurosciences, biotechnologie, physique, informatique etc., comme vu dans le paragraphe *1.2.4.5 Military Brain Science* (Jin et al., 2018 ; Brunyé et al., 2020). Elle vise à mieux comprendre le fonctionnement du cerveau, à perturber celui de l'adversaire par des technologies comme les micro-ondes ou les drogues, et à renforcer les capacités cognitives des troupes chinoises (Beauchamp-Mustafaga, 2023 ; Charon & Jeangène Vilmer, 2021). La modélisation cognitive, qui combine sciences cognitives et IA, vise à simuler les processus mentaux pour anticiper les réactions des cibles et optimiser l'efficacité des opérations psychologiques (Hung & Hung, 2022 ; Charon & Jeangène Vilmer, 2021 ; Yeh, 2021).

1.7.2.3 Politique de réseaux sociaux

La politique de réseaux sociaux de la Chine combine censure interne et influence externe via ses plateformes numériques. Son système de contrôle de l'information, le *« Grand Pare-feu »*, bloque l'accès à de nombreux sites étrangers comme Facebook, Twitter ou la BBC afin de restreindre les contenus jugés sensibles et maintenir la stabilité sociale (Zucchi, 2022).

Ainsi, TikTok et Douyin, tous deux produits par l'entreprise chinoise ByteDance, fonctionnent différemment. Douyin, réservé au marché chinois, est soumis à une censure stricte, tandis que TikTok est parfois accusé de servir les intérêts du Parti communiste chinois en diffusant des contenus alignés sur sa propagande et en censurant des sujets sensibles pour la Chine tels que les droits des Ouïghours (Robin, 2022 ; Marpaung, 2025) (voir paragraphe *1.6.3.1 Offensif*). Considéré par certains experts comme un vecteur potentiel de guerre cognitive, TikTok a été interdit dans plusieurs pays pour des raisons de sécurité nationale, à commencer par l'Inde en 2020 (Košková, 2025 ; Robin, 2022).

1.7.3 Moyen-Orient

Hanson (2004) estime que certains groupes islamistes radicaux du Moyen-Orient exploitent leur connaissance des sociétés occidentales, en s'appuyant sur les failles perçues de leur culture, sur des images et une rhétorique conçue pour *« altérer l'esprit et éroder la volonté [de l'Occident] »*. Selon lui, la bataille idéologique se poursuit même lorsque la confrontation militaire est perdue, notamment par l'exploitation d'images et de récits conçus pour semer le doute, la culpabilité ou la division.

Al-Qaïda utilise internet comme vecteur de manipulation idéologique pour diffuser une vision radicale à travers des récits émotionnels, religieux ou polarisants. Les contenus visent à susciter l'adhésion idéologique, à renforcer les tensions identitaires et à recruter des individus vulnérables, en situation de détresse, d'isolement ou en quête de sens. Ce discours, adapté aux contextes sociopolitiques, constitue une forme de propagande visant à radicaliser et fidéliser durablement les recrues dans la communauté jihadiste (Borgonovo, 2022 ; Soygeniş, 2014 ; Badawy & Ferrera, 2018 ; Green, 2008).

1.7.4 Syndrome de La Havane

Le syndrome de La Havane désigne un ensemble de symptômes apparus à partir de 2016 chez des diplomates, militaires et agents de renseignement américains et canadiens. Ces troubles incluent des acouphènes, vertiges, pertes de mémoire, troubles visuels ou cognitifs (Connolly et al., 2024 ; McCreight, 2022). Un rapport du Bureau du directeur du Renseignement national américain (2022) conclut qu'aucune cause certaine n'est identifiée, mais que certains cas pourraient résulter de stimuli externes, tels que des radiofréquences ou ultrasons. Les facteurs psychosociaux et environnementaux n'expliquent pas seuls les symptômes (Office of the Director of National Intelligence, 2022). Si l'origine s'avérait intentionnelle, ces effets pourraient relever d'une attaque utilisant les NBIC dans une logique de guerre cognitive, ciblant des individus-clés pour les perturber ou les affaiblir.

1.8 Enjeux

1.8.1 Les dangers de la guerre cognitive

La guerre cognitive menée contre nos démocraties représente un danger évident : comme nous l'avons vu lors des chapitres précédents, elle peut inclure l'influence des élections démocratiques, la manipulation de l'opinion des populations, la baisse de l'attention et de l'intérêt de ces dernières pour les études avancées, la baisse en conséquence des capacités d'innovation, la compartimentation et la polarisation des opinions produisant d'importantes frictions sociales… Ce sont tout un ensemble de conséquences qui peuvent contribuer à déstabiliser et fragiliser différents aspects et acteurs d'une nation.

1.8.2 Vulnérabilités des sociétés démocratiques et non-démocratiques

Les sociétés démocratiques sont particulièrement vulnérables aux stratégies de guerre cognitive, car l'information y circule librement (Green, 2008). Cette ouverture facilite l'accès à des données sur les comportements, les habitudes et les fragilités des populations, que des acteurs malveillants peuvent exploiter pour produire des contenus d'influence ciblés (Buchler, 2021). Il est aujourd'hui facile et peu coûteux de diffuser des informations à large échelle sur les médias sociaux et, d'après Wunder (2021), les

plateformes ont peu d'intérêt à réguler ces comptes d'influence, qui génèrent beaucoup de contenu et d'engagement. Les utilisateurs eux-mêmes montrent peu d'intérêt pour la vérification des informations, qui demande du temps, de l'attention et des efforts, rendant plus facile la consommation de contenus rapides, émotionnels et compatibles avec leurs opinions. L'utilisateur tend alors à devenir passif face aux contenus suggérés par les algorithmes, ce qui soulève la question du contrôle qu'exercent ces derniers sur sa consommation d'informations et ses connaissances. Pour Holeindre (2017), une des faiblesses des démocraties viendrait de la culture *« chevaleresque »* qui mènerait à *« cette arrogance de la force et ce mépris de la ruse »*, et potentiellement à une ignorance voire un déni de l'existence de la guerre cognitive.

Par ailleurs, la méfiance envers l'État facilite l'acceptation de discours conspirationnistes qui fragilisent la cohésion sociale (Orinx & Struye de Swielande, 2021). Cette méfiance peut également se manifester dans les états non-démocratiques. Ces derniers, souvent caractérisés par une forte concentration des médias et un contrôle étatique de l'information, sont confrontés à d'autres types de risques informationnels, tels que la propagation de rumeurs souterraines et déstabilisatrices sur des canaux alternatifs difficiles à contrôler (Roberts, 2018).

1.8.3 Facteurs culturels facilitant l'utilisation de la guerre cognitive

Les régimes autoritaires présentent souvent une résilience apparente face aux stratégies de guerre cognitive, en raison d'un contrôle renforcé de l'information et des représentations collectives. L'accès à l'information y est fortement régulé, et les opinions de la population peuvent être orientées ou cadrées par le pouvoir politique. Par exemple, Shtepa (2021) affirme que le gouvernement russe mène une guerre cognitive contre sa propre population pour renforcer ses fondations idéologiques avec des *« liens spirituels »* ou *« valeurs russes traditionnelles »* (voir paragraphe *1.7.1 Russie*).

À l'inverse, les sociétés démocratiques, plus ouvertes à la pluralité des discours et à la libre circulation de l'information, se révèlent plus vulnérables aux attaques cognitives, notamment lorsque leurs récits collectifs et symboliques sont affaiblis. Comme le souligne Valette (2024), la guerre cognitive agit aussi sur le plan sémiotique et culturel, en s'attaquant aux créations symboliques qui fondent la légitimité et la cohésion des sociétés, notamment le *récit national*, *« pièce maîtresse d'une défense cognitive à inventer »*.

La culture et les valeurs d'un État peuvent également favoriser une approche offensive de la guerre cognitive : certaines traditions stratégiques acceptent davantage les *« zones grises »*, avec moins de restrictions éthiques. La culture des « trois guerres » chinoises (voir le paragraphe *1.7.2.1 Culture de guerre chinoise*) en est un exemple. Par ailleurs, la Chine a *« une approche du temps très différente de celle de l'Occident »*, favorisant les interventions avec des objectifs sur le temps long (Orinx & Struye de Swielande, 2021 ; Charon & Jeangène Vilmer, 2021).

1.8.4 Stratégies défensives et posture proactive

Face aux dangers de la guerre cognitive, exacerbés par l'utilisation massive des réseaux sociaux (62% des français s'informent aujourd'hui via les réseaux sociaux – Patino, 2023) et par l'immédiateté du partage de l'information, il paraît important de développer des stratégies pour s'en protéger.

Plusieurs solutions ont été mentionnées pour contrer des actions spécifiques de guerre cognitive. Parmi les contre-mesures identifiées pour lutter contre l'influence du public et les fausses informations sur les

médias sociaux, nous pouvons compter l'éducation du public, la sensibilisation à la désinformation, la modération automatique et humaine, la diffusion de contre-informations, ainsi que l'instauration de réglementations juridiques (Backes & Swab, 2019 ; Cao et al., 2021 ; Wunder, 2021 ; Bertrand, 2023b).

Ducourneau (2024) suggère la mise en place d'un *« design lab »* dédié à la sécurité cognitive. Il permettrait de travailler sur la résilience face à la manipulation et à l'ingérence dans les processus de pensée individuels et collectifs, en s'appuyant sur les sciences sociales et la technologie. L'objectif serait de développer des outils robustes pour comprendre, former et répondre aux menaces cognitives, grâce à un processus itératif d'intégration des résultats pour imaginer une solution adaptée à un contexte opérationnel dynamique.

Une autre approche de protection et de prévention consiste à utiliser la guerre cognitive de manière défensive : *« il faut être en mesure de protéger nos cerveaux tout autant que d'améliorer les capacités de compréhension et de décision de nos cerveaux »* (Autellet, 2021). Les outils de guerre cognitive peuvent être employés pour sensibiliser et éduquer les populations via les médias et les réseaux sociaux (Battrawi & Muhtaseb, 2013), améliorer la préparation cognitive (Morrison & Fletcher, 2002), ou même augmenter la cognition des soldats (Jin et al., 2018). Nous pourrions également envisager des outils d'aide à la décision prenant en compte les biais cognitifs et les potentielles agressions de guerre cognitive.

La première étape pour organiser une solution globale consiste à analyser l'adversaire, comprendre ses stratégies de guerre cognitive ainsi que les outils et tactiques qu'il utilise (Bebber, 2024). Cela permettrait aux cibles de détecter rapidement les offensives cognitives et dissiper le brouillard de la guerre (Weldon, 2021). Il est nécessaire de poursuivre les recherches sur ce sujet afin d'élaborer des solutions à la fois systémiques et systématiques.

1.8.5 Agir dans le champ de la guerre cognitive

Dans un monde qui change, Stoianov (2021) prédit que *« l'esprit humain va devenir la cible prioritaire »*. Dans ce contexte, aucun pays ne devrait négliger l'utilisation de la guerre cognitive pour défendre ses intérêts et se contenter de s'en protéger (Chauvancy, 2021).

Giordano (2018) plaide pour un engagement total dans la guerre cognitive et un investissement important dans la recherche en sciences cognitives et neurologiques. Il affirme que la Chine et la Russie ont de l'avance sur le sujet, disposant notamment d'infrastructures de recherche médicale dont les résultats peuvent être transférés au domaine militaire.

En 2021, le général Burkhard, chef d'État-Major des Armées françaises, évoquait la nécessité de *« gagner la guerre avant la guerre »* et appelait à : *« promouvoir un état d'esprit qui permette de gagner la bataille des idées »*. Cela témoigne d'une reconnaissance croissante de l'importance de la guerre cognitive. Un autre signe de cette reconnaissance est la proposition étudiée par l'OTAN de la faire entrer parmi ses domaines d'opération, aux côtés de la terre, la mer, l'air, l'espace et le cyberespace, pour agir dans les dimensions humaines et cognitives (Le Guyader & Cole, 2020 ; Montocchio, 2021). Par ailleurs, en 2022,

l'Agence de l'Innovation de Défense a lancé un appel à projets pour soutenir la recherche et l'innovation sur la thématique de la guerre cognitive[2].

1.9 Limites et critiques de la guerre cognitive

Certaines actions de déstabilisation ou d'influence peuvent être détectées et contrées. Par exemple, une entreprise utilisant une stratégie de guerre cognitive contre ses concurrents ou ses clients pourrait être exposée par une fuite de données ou un lanceur d'alerte, ce qui nuirait à son image publique. L'épisode du charnier de Gossi, mis en scène par le groupe Wagner en 2022 au Mali mais dévoilé par l'armée française dans une opération de contre-information, en est un exemple (Mwai, 2022). Les tentatives de démoralisation et de sape de la confiance ont même parfois l'effet inverse à celui recherché (Wilde, 2024), et ces stratégies ne suffisent pas toujours pour gagner ou éviter une guerre (Takagi, 2022).

Par ailleurs, la guerre cognitive, visant à influencer la pensée et la prise de décision de personnes à leur insu, soulève des défis éthiques. À cet égard, elle peut être comparée aux *« nudges »*, ces incitations subtiles intégrées à l'environnement de décision qui orientent les choix des individus sans restreindre leur liberté (Loewenstein & Chater, 2017). Il s'agit également d'un outil de manipulation, mais c'est son utilisation qui détermine ses caractéristiques éthiques. Par exemple, un nudge poussant un consommateur à acheter plus ou à un prix plus élevé sera jugé négativement, tandis qu'un nudge encourageant un comportement plus respectueux, comme l'autocollant en forme de mouche dans les urinoirs pour améliorer la précision et diminuer les éclaboussures (Evans-Pritchard, 2013), sera perçu positivement. Il en va de même pour la guerre cognitive, qui doit être utilisée de manière raisonnable et justifiable.

➢ Bilan des paragraphes 1.7 à 1.9 : les points-clés

Nous avons vu que certains pays ont une culture qui accepte mieux les *« zones grises »*, et pour lesquels l'utilisation de la guerre cognitive est plus naturelle. La Chine et la Russie sont largement citées comme exemples dans la littérature.

Les démocraties sont particulièrement vulnérables aux actions de guerre cognitive car ouvertes à l'information, à la diversité d'opinion et au numérique. La guerre cognitive vise à affaiblir la cohésion sociale, la confiance dans les institutions, et les capacités d'innovation.

Cependant, elle peut échouer si elle est détectée et contrée correctement. Elle soulève des enjeux éthiques majeurs et peut produire des effets inverses à ceux recherchés si mal calibrée.

La guerre cognitive cible la pensée, les émotions et mécanismes mentaux pour influencer ou entraver les décisions. Pour mieux comprendre comment ces attaques opèrent et pourquoi elles sont efficaces, il convient de s'intéresser aux vulnérabilités de la décision humaine. Le *Chapitre 2* explore les processus de prise de décision, leurs fragilités et les moyens par lesquels elle peut être influencée ou orientée.

[2] https://anr.fr/fr/detail/call/accompagnement-specifique-des-travaux-de-recherches-et-dinnovation-defense-appel-thematique-sur-l-1/

Chapitre 2 - Décision, biais et influence de la décision

2.1 Qu'est-ce que la décision ?

Dans le cadre de cette thèse, nous nous intéressons à la prise de décision de personnes en situation d'autorité, en particulier à celle des décideurs militaires, dans un cadre de gestion de crise liée à la guerre cognitive.

2.1.1 Définition

Allain (2013) définit la décision comme le *« fait d'effectuer un choix entre plusieurs modalités d'actions possibles lors de la confrontation à un problème, le but étant de le résoudre en traduisant le choix fait en un comportement »*. La décision comble une rupture de causalité. Elle répond à la nécessité d'un choix non évident entre plusieurs options ou actions dont l'issue est incertaine. En effet, si le choix est évident et la décision totalement rationnelle, il ne s'agit plus d'une décision mais d'un processus : il n'existe pas de décision automatique (Chaudron et al., 1993 ; Bartheye & Chaudron, 2018). Il s'agit de l'une des facultés indispensables à l'être humain : faire des choix *« allant dans le sens de nos intérêts propres »* est *« indispensable à l'adaptation et à l'autonomie »* (Allain, 2013). Mais ce choix est aussi l'interstice dans lequel peuvent s'immiscer l'erreur, le biais ou la manipulation.

Les humains prennent de nombreuses décisions tous les jours ; en plus des choix de la vie quotidienne et des grandes décisions prises dans différents domaines (famille, travail, finances…), *« de nombreux jugements se résument naturellement à des choix binaires, car toutes sortes de questions peuvent être formulées comme des demandes d'estimation de l'événement en question et de son complément (événement ou non-événement) »* (Hilbert, 2012). Chacun de ces choix se réfère à des enjeux et conséquences plus ou moins importants : gain estimé, risque… Ces décisions peuvent également être prises suivant la maîtrise du résultat attendu : résultat certain, résultat incertain avec connaissance du taux de risque ou chances de succès, résultat incertain avec non connaissance des chances de succès (Allain, 2013 ; Lemaire, 1999).

> ➢ **Point-clé : définition de la Décision**
>
> Décision : *« Choix conscient et raisonné suivi d'une action aboutissant à une modification de l'individu, de l'environnement et de leur relation »* (Albou et al., 1990).

2.1.2 Les étapes de la prise de décision

Si en français la décision est *« prise »*, en anglais elle est *« construite »* (decision making). Cette notion anglophone nous rappelle qu'une décision n'est pas seulement un choix entre plusieurs options, elle est le résultat d'un processus qui comprend la récolte d'informations, l'analyse et l'interprétation de la situation, la considération de différentes solutions (élaboration et évaluation de différentes hypothèses, leurs risques et leurs gains et pertes potentiels), puis le choix d'une solution et enfin sa mise en œuvre

(Huber et al., 2011 ; Allain, 2013 ; Lemaire, 1999). Un exemple de processus décisionnel formalisé largement repris est la boucle OODA en quatre étapes : Observer, Orienter, Décider, Agir (Orr, 1983).

La construction d'une conscience de la situation adaptée est une étape indispensable à la construction d'une décision. Celle-ci dépend des informations disponibles et des connaissances préalables du décideur, mais aussi de ses idées préconçues ou attentes, ses expériences antérieures, son état cognitif… Une conscience de situation correcte ne garantit cependant pas la sélection d'une solution adéquate : cette dernière peut être mal choisie à cause d'un processus décisionnel inadapté, un manque d'expérience ou d'entraînement, des contraintes organisationnelles etc. La personnalité et autres facteurs individuels du décideur interviennent aussi dans la prise de décision. De même, une décision correcte ne garantit pas une exécution optimale (Endsley & Garland, 2000).

Par ailleurs, Coccia (2020) nous rappelle qu'une décision est souvent composée de plusieurs sous-décisions : une solution peut engendrer d'autres problèmes auxquels il faudra trouver d'autres solutions. Il propose ainsi la construction d'un arbre de décision incluant les conséquences de chaque décision, qui permettra alors de sélectionner le plus objectivement possible la suite de décisions qui aura le meilleur résultat final.

Après la décision, le décideur peut ressentir du regret si le résultat n'est pas satisfaisant, ou parfois si le nombre d'options disponibles était élevé ou si la décision est réversible. Une autre réaction possible à une issue décevante est de restructurer son expérience et sa pensée, trouvant des raisons qui expliquent pourquoi c'était finalement une décision adaptée : c'est ainsi qu'une décision peut apparaître meilleure à celui qui l'a prise si elle est irréversible, par rapport à la même décision et au même résultat, mais réversible. En un mot, les individus ont tendance à justifier leur choix et à mieux s'en contenter lorsqu'il est trop tard pour en changer : c'est la rationalisation *a posteriori* (Festinger, 1957 ; Dietrich, 2010 ; Botti & Iyengar, 2004).

2.1.3 La construction de la conscience de situation

Si l'on considère la décision comme un choix conscient et non automatique qui nécessite une évaluation des résultats potentiels, alors construire une conscience ou représentation de la situation adaptée est indispensable. En effet, plus l'information est imparfaite (imprécision, inconsistance, incertitude, dimensions, hétérogénéité, complexité…), plus la prise de décision est difficile ; atteindre la supériorité informationnelle est donc primordial (Sedogbo et al., 2007).

La conscience de situation peut être individuelle ou collective. Deux personnes face à la même situation peuvent en avoir une représentation différente. Le partage de la conscience de situation constitue dès lors un enjeu majeur dans les environnements collaboratifs, qu'ils impliquent des interactions humain-humain, humain-machine ou inter-systèmes. Endsley (1995, 2015) souligne que la conscience de situation partagée est essentielle à la performance collective, mais difficile à atteindre en raison des différences de perception, d'attention et de modèles mentaux entre acteurs.

Pour Lebraty (2004), *« comprendre une situation signifie que le décideur accepte un niveau de représentation qu'il juge suffisant »* : la conscience de situation *« n'est pas une propriété de la situation…, elle n'existe que par rapport à une intention de l'opérateur »* (Amalberti, 1996). Cela montre que la conscience de situation doit tenir en une représentation synthétique : il n'est pas nécessaire de connaître et comprendre toutes les informations possibles concernant la situation, cela noierait l'information utile

à la décision ou la tâche associée ; surtout que *« nous sommes incapables de traiter l'ensemble des informations en provenance de notre environnement »* (Allain, 2013). Il faut donc extraire et regrouper (*« information fusion »*) l'information adéquate en fonction de l'objectif (événements passés et informations pertinentes, parties prenantes, risques, anticipation des conséquences, etc.), afin de se construire une représentation juste et synthétique de la situation (Sedogbo et al., 2007).

L'un des moyens les plus évidents pour un décideur est de raisonner en boucles courtes pour tester ses solutions et faire évoluer sa représentation de la situation en fonction des résultats (Lebraty, 2004). Pour construire une conscience de la situation efficace, il faut percevoir correctement les informations pertinentes, les interpréter à l'aide d'un modèle mental adapté, et pouvoir anticiper l'évolution de la situation en s'appuyant sur une mémoire et une attention suffisantes. Il faut également éviter la surcharge informationnelle en se concentrant sur l'information utile à la tâche. Une représentation pertinente repose sur l'identification des éléments pertinents pour la tâche ou la prise de décision, de leurs structures spatiales, temporelles et organisationnelles, ainsi que des actions envisageables. Enfin, il est important de fusionner les informations issues de sources diverses (capteurs, bases de données, observations humaines…) pour obtenir une vision plus cohérente et fiable que celle qu'offrirait une source isolée (Laudy et al., 2006 ; Yi et al., 2025).

2.1.4 Les facteurs de la prise de décision

Qu'est-ce qui fait que l'on prend une décision plutôt qu'une autre ?

Caractéristiques propres à la décision : D'après Elbanna et Child (2007), les principales caractéristiques de la décision à prendre sont l'importance que l'on y attache, l'incertitude (risque) qui y est liée, et le motif ou la motivation de la décision. Il faut y ajouter le coût associé aux différentes solutions envisagées et les performances attendues de chaque solution.

Caractéristiques propres à l'environnement : Outre le contexte et la situation qui fait l'objet d'une décision, Elbanna et Child (2007) avancent que la décision dépend de l'incertitude liée à l'environnement (par exemple, pour une entreprise, il faut prendre en compte les incertitudes du secteur comme les technologies ou les compétiteurs) et le niveau d'hostilité de l'environnement.

Caractéristiques propres au décideur : Le style cognitif du décideur est un facteur important dans la prise de décision et l'évaluation du risque (Engin & Vetschera, 2017). Par exemple, une expérimentation de Henderson et Nutt (1980) a montré que des décideurs expérimentés ayant une aversion au risque sont réticents à adopter des projets d'expansion de capacité d'entreprises ou d'hôpitaux. Au contraire, les personnalités tolérantes au risque sont plus susceptibles d'adopter les mêmes projets. Dietrich (2010) cite comme facteurs individuels d'influence de la décision : les expériences passées (si une solution a bien fonctionné pour une personne par le passé, elle aura tendance à la reproduire dans une situation similaire), l'âge et le statut socioéconomique, la confiance en soi et les capacités cognitives.

2.1.5 Les niveaux de contrôle

La qualité de la prise de décision et de la solution adoptée tout comme la conscience de situation, dépendent en partie du niveau de contrôle de l'opérateur ou du décideur sur la situation et la tâche à

accomplir. Différents niveaux de contrôle ont été décrits par Rasmussen 1986 puis par Hollnagel en 1993 (Lebraty, 2004). Nous proposons de les mettre en correspondance dans le Tableau 9.

Hollnagel propose quatre niveaux de contrôle : le niveau d'alerte, où le décideur a une compréhension inadaptée de la situation et fait face à une pression temporelle ; le niveau opportuniste, où le décideur n'a pas une compréhension complète de la situation, anticipe peu et utilise des heuristiques ; le niveau tactique, où le décideur utilise des règles et protocoles, et la pression temporelle moyenne permet une anticipation à moyen terme ; et le niveau stratégique, où le décideur a une compréhension complète de la situation et dispose de temps pour anticiper à long terme et planifier (Hollnagel, 1993).

Rasmussen distingue trois niveaux de contrôle : le niveau basé sur des connaissances, qui est le moins automatisé car le décideur, par exemple peu familier avec la situation, doit l'analyser en profondeur ; le niveau basé sur des règles, où le décideur utilise des règles qu'il a mémorisées et s'est appropriées ; et le niveau d'expertise le plus élevé, dans lequel le décideur utilise de nombreux automatismes et la tâche lui demande un faible coût cognitif (Rasmussen, 1986 ; Lebraty, 2004).

Tableau 9 : Les niveaux de contrôle de Hollnagel et Rasmussen (Lebraty, 2004)

Les niveaux de contrôle d'après Hollnagel (1993) et Rasmussen (1986)				
	Non contrôle	**Faible contrôle**	**Bon contrôle**	**Contrôle élevé**
Hollnagel	Alerte	Opportuniste	Tactique	Stratégique
Rasmussen		Niveau basé sur des connaissances	Niveau basé sur des règles	Niveau d'expertise élevée

Ces deux modèles montrent que la qualité de la décision dépend à la fois de l'expertise du décideur sur la situation rencontrée, de la complexité de celle-ci, ainsi que de la pression temporelle à laquelle il est soumis. Toutefois, même dans des conditions de contrôle élevé, la décision humaine reste soumise à des limites intrinsèques. Il convient donc d'examiner plus en détail les fragilités et les menaces qui peuvent affecter la qualité du processus décisionnel.

➢ Bilan du paragraphe 2.1 : les points-clés

Une décision n'est pas un automatisme, mais un choix conscient entre plusieurs options, fait dans un contexte incertain et en fonction d'un objectif. Elle consiste en plusieurs étapes souvent itératives et non séquentielles : perception de la situation, élaboration d'options, évaluation, choix, mise en œuvre.

La qualité de la décision dépend de la conscience de la situation construite à partir des informations perçues, des connaissances et du contexte. Plusieurs facteurs influencent ce processus : caractéristiques du décideur (expérience, confiance, style cognitif), de l'environnement (pression, incertitude) et de la tâche (enjeux, contraintes).

Différents niveaux de contrôle permettent de caractériser la manière dont un individu maîtrise ou subit la situation décisionnelle.

2.2 Quelles sont les fragilités et les menaces pour la décision ?

Si la prise de décision vise à trouver une solution adaptée et satisfaisante, la théorie de la rationalité limitée prédit que les limitations computationnelles et cognitives humaines (mémoire, connaissances insuffisantes, perception incomplète de la situation, déficience d'analyse…) face à une réalité complexe font que la solution choisie est rarement optimale (Simon, 1979 ; Bhui et al., 2021). Nous explorons ici certains éléments qui peuvent menacer la recherche d'une solution satisfaisante au regard de ces limitations.

2.2.1 Les biais cognitifs

Pour comprendre le concept de biais cognitif, il faut d'abord connaître le concept d'heuristique. Lorsqu'un individu doit réagir rapidement à une situation, il mobilise parfois un mode de raisonnement intuitif et automatique, fondé sur l'expérience et les émotions, plutôt qu'un raisonnement délibératif et analytique plus lent et coûteux en ressources cognitives (Evans & Stanovich, 2013). Ces raisonnements intuitifs s'appuient sur des mécanismes simplifiés de traitement de l'information, appelés heuristiques.

Les heuristiques sont définies par Shah et Oppenheimer (2008) comme des stratégies générales de prise de décision qui sont basées sur peu d'informations, mais qui sont très souvent correctes ; il s'agit donc de raccourcis mentaux qui réduisent la charge cognitive associée à la prise de décision. Les auteurs listent quatre heuristiques importantes qui sont : l'heuristique de représentativité, l'heuristique de disponibilité, l'heuristique d'ancrage et l'heuristique d'ajustement. Ces heuristiques sont utiles car elles donnent un résultat correct dans la majorité des cas ; mais elles peuvent aussi mener à des erreurs de jugement, qui sont appelées des biais cognitifs. Un exemple simple est donné par Frederick (2005) avec le problème de la batte et de la balle : une question courte à laquelle il faut tenter de répondre rapidement.

> *Une batte et une balle coûtent au total 1,10 $. La batte coûte 1,00 $ de plus que la balle. Combien coûte la balle ?*

À ce problème, la réponse donnée par plus de 50% des étudiants de Princeton et de l'Université de Michigan était incorrecte : 10 centimes (la réponse juste étant 5 centimes). Cette expérimentation montre que de nombreuses personnes trouvent une réponse erronée car elles utilisent des raccourcis de pensée ou heuristiques, qui leur permettent dans la majorité des cas de trouver la bonne réponse intuitivement. L'utilisation de ces heuristiques peuvt être encouragée par la pression temporelle (*« donner une réponse rapidement »*). Or, pour trouver ici la bonne réponse, il faut faire l'effort mental de poser le calcul.

Les biais cognitifs sont un écart systématique (c'est-à-dire non aléatoire et donc prévisible) par rapport à la rationalité dans le jugement ou la prise de décision (Blanco, 2017). En effet, les êtres humains et leurs choix sont *« prévisiblement irrationnels »* (*« predictably irrational »*), comme l'indique Dan Ariely en 2008. Les biais cognitifs nous dirigent toujours vers les mêmes types de déviations par rapport à ce que la théorie classique des probabilités prédit comme étant le résultat optimal de nos choix. Si les erreurs sur les choix étaient aléatoires, elles se rattraperaient et la moyenne serait juste ; or, ce n'est pas le cas (Hilbert, 2012). Cela entraîne l'apparition de déviations systématiques dont héritent souvent les IA génératives, qui s'inspirent des productions humaines (Martínez et al., 2022 ; Osoba & Welser, 2017). Pour comprendre ce principe, nous pouvons prendre l'exemple des illusions d'optique : même si une personne sait qu'il s'agit d'un leurre, elle ne peut s'empêcher d'en être victime et sa perception de la réalité est systématiquement erronée (Blanco, 2017).

Les causes des biais cognitifs sont multiples. L'architecture biologique qui constitue l'humain le limite : nos capteurs (visuels, auditifs, sensoriels), nos ressources cognitives, notre mémoire ou encore notre attention, restreignent la quantité d'informations que nous pouvons traiter correctement en simultané. La motivation et les émotions influencent notre perception et notre prise de décision, pour laquelle elles sont nécessaires. L'influence sociale atténue certains biais et en renforce d'autres, comme dans le cas du biais de conformité au groupe. Enfin, notre recours spontané à des heuristiques ou raccourcis mentaux, utiles pour alléger la charge cognitive, peut entraîner des erreurs de jugement systématiques (Blanco, 2017).

De nombreux biais cognitifs ont déjà été identifiés et la recherche sur le sujet est en constante évolution, de nouveaux biais étant découverts régulièrement. Baron en listait déjà 53 en 2007, et certains inventaires en recensent aujourd'hui plusieurs centaines, bien que le manque d'uniformité dans la terminologie complique leur recensement : un même biais peut porter plusieurs noms, et un même nom désigner des biais différents (Hilbert, 2012). Dans le paragraphe suivant, nous nous intéressons spécifiquement aux biais liés à la prise de décision.

2.2.2 Les biais de la décision

De nombreux biais cognitifs ou failles cognitives peuvent être associés aux différentes étapes de la décision : lors de la prise d'informations, de son traitement par le cerveau humain, lors du choix d'une solution, etc. (Arnott, 1998). Ainsi, ils peuvent être utilisés pour pousser une personne à prendre des raccourcis de pensée qui vont influencer sa décision, par exemple en lui présentant des informations sous une certaine forme qui fait appel aux biais cognitifs ; ou encore pour ralentir ou empêcher la prise de décision. D'après Dietrich (2010), les biais cognitifs amènent les individus à *« accorder plus de crédit aux observations attendues et aux connaissances antérieures, tout en rejetant les informations ou les observations perçues comme incertaines »* ; Lebraty (2004) indique que *« les biais peuvent contribuer à faire dévier le décideur de son intention, mais, facteur aggravant, les biais masquent cette déviation »*. Il est donc important d'identifier ces biais de la décision, afin de savoir ce qui peut être utilisé et ce dont il faut se protéger. Par exemple, un comité d'experts des National Academies of Sciences, Engineering, and Medicine (Endsley et al., 2022) estime qu'il serait pertinent de créer des assistants intelligents pour réduire les biais de la décision. Cependant, d'après Mercier (2017), nous aurions plutôt tendance à nous tromper dans le bon sens, et un défaut d'objectivité donne parfois de meilleures décisions.

L'un des biais de la décision les plus connus est le biais d'autorité *(authority bias)*, illustré en 1961 par Milgram : une personne soumise à des ordres peut se retrouver déresponsabilisée, certains allant jusqu'à *« tuer »* un sujet en lui infligeant des chocs électriques létaux au nom de la science, sur ordre d'un expérimentateur ; le sujet est en réalité un acteur et ne reçoit aucun choc électrique (Milgram, 1963). Cependant, il existe de nombreux autres biais en lien avec la prise de décision, dont nous donnerons quelques exemples ici.

Arnott (1998) propose une taxonomie des biais de la décision les rangeant dans 6 catégories :

- Les biais liés à la mémoire : biais rétrospectif (tendance à croire, après coup, qu'un événement était prévisible – Dietrich, 2010), biais de recherche en mémoire (tendance à se remémorer l'information de manière orientée ou sélective), biais de similarité (tendance à se souvenir d'événements similaires comme plus fréquents qu'ils ne le sont réellement), etc.

- Les biais statistiques : biais de taux de base (tendance à considérer des informations anecdotiques plutôt que les statistiques générales), biais de corrélation illusoire (tendance à surestimer ou imaginer une corrélation entre deux événements), biais d'échantillonnage (tendance à tirer des conclusions d'un échantillon non représentatif), etc.
- Les biais de confiance : biais de complétude (tendance à croire qu'une information est plus fiable parce qu'elle semble complète ou détaillée), biais de contrôle (surestimation de sa capacité à contrôler des événements qui sont en réalité dus au hasard), biais de confirmation (tendance à observer ce que l'on attend des observations et à ainsi justifier des choix irrationnels – Dietrich, 2010)… Dietrich mentionne également le biais de croyance (belief bias), soit la dépendance excessive à l'égard des connaissances préalables pour la prise de décision, qui pourrait être considéré comme un biais de confiance.
- Les biais d'ajustement : biais d'ancrage et d'ajustement (tendance à se fier excessivement à la première information reçue), biais de conservatisme (tendance à mettre à jour ses croyances trop lentement face à de nouvelles preuves), etc.
- Les biais de présentation : biais de cadrage (tendance à réagir différemment selon la manière dont une information est formulée), biais de mode de présentation (impact du format ou du canal de diffusion d'une information sur sa perception), biais d'ordre (influence de l'ordre dans lequel les informations sont présentées), biais d'échelle (influence de la manière dont une échelle est représentée dans une visualisation), etc.
- Les biais situationnels : biais d'atténuation (tendance à sous-estimer la gravité d'un avertissement, surtout s'il semble lointain, abstrait ou incertain), biais d'escalade d'engagement et des résultats irrécupérables (tendance à continuer à investir dans une décision perdante parce que l'on y a déjà consacré du temps, de l'argent ou de l'énergie) (Dietrich, 2010), biais d'habitude (tendance à agir selon des habitudes passées, même si elles ne sont plus adaptées à la situation actuelle), etc.

Concernant le biais de corrélation illusoire, Hilbert (2012) donne l'exemple *« délicat »* des décideurs qui *« forment de fausses associations entre des personnes ayant des comportements rares (typiquement négatifs) et l'appartenance à des groupes statistiques minoritaires »*.

« L'Elaboration Likelihood Model », ou théorie de la probabilité d'élaboration, nous apprend que lorsqu'une personne est intéressée, motivée et concentrée sur le message, elle traite l'information plus en profondeur, ce qui rend les changements d'opinion plus résistants aux contre-arguments et les changements de comportement plus durables dans le temps qu'en cas de traitement superficiel de l'information. Ce modèle repose sur plusieurs études quantitatives (Petty & Cacioppo, 1984). Nous pouvons en déduire qu'une personne dans cette situation aurait tendance à être moins sujette à certains biais comme les biais de présentation et d'ajustement ; mais le principe même des biais cognitifs est qu'ils sont inhérents à la cognition humaine et inconscients, il est donc très difficile de les éviter.

2.2.3 Les manipulations et influences de la décision

Si la guerre cognitive vise prioritairement les décisions stratégiques, étudier les techniques de manipulation en contexte quotidien permet de mieux comprendre les failles cognitives universelles, que l'on retrouve dans les environnements décisionnels critiques. Ainsi, lorsque le sujet de la manipulation est abordé, certains évoqueront le célèbre *« Petit traité de manipulation à l'usage des honnêtes gens »* de Joule et Beauvois (1987), un essai de psychologie sociale qui revient sur diverses techniques de

manipulation connues et accessibles à toute personne informée (pied dans la porte, porte au nez…). Ils mentionnent notamment des techniques de manipulation qui peuvent être utilisées dans la sphère privée pour aider à convaincre ou persuader, mais aussi des techniques commerciales qui peuvent aider à vendre en séduisant et manipulant le consommateur sans qu'il en ait nécessairement conscience (Moran, 2020).

Colon (2021) définit la manipulation comme *« l'art de fausser la réalité et d'influencer les individus à leur insu »*. Elle est souvent associée à la persuasion, qui consiste à *« agir en douceur sur les conduites des individus ».* Heilbrunn (2022) énonce également : *« Qu'ils passent par la séduction ou par la pression, nous sommes tous soumis quotidiennement à des jeux d'influence. La manipulation commence lorsque nous n'en avons plus conscience »*. Colon (2021) rappelle, comme Joule et Beauvois, que la manipulation peut être positive, comme lorsqu'il s'agit de manipuler un fumeur pour qu'il cesse de fumer. Gass & Seiter (2022) la comparent à l'apprentissage des arts martiaux : comme la guerre cognitive, il s'agit d'un outil, et il appartient à ceux qui la pratiquent d'en avoir un usage responsable.

Il est important de distinguer la propagande de la persuasion. La propagande est plus proche du marketing politique, où l'objectif est de manipuler l'opinion publique, tandis que la persuasion cherche à convaincre par des arguments rationnels. Le lavage de cerveau, bien que souvent évoqué, n'est pas un concept scientifique. Il a été utilisé par la propagande américaine pour renforcer l'attitude négative des Américains envers les pays adversaires (communistes) (Dubois, 1997). Cependant, en ligne, les possibilités et opportunités de manipuler ou d'être manipulé explosent, et encore plus avec les réseaux sociaux *(*voir paragraphe *1.6.3 Réseaux sociaux)* et les intelligences artificielles génératives. Huang et Wang (2023) ont conclu de leur méta-analyse de 121 études empiriques, que les intelligences artificielles étaient aussi persuasives que des humains, tandis que Costello et al. (2024) ont montré qu'un dialogue avec une IA générative pouvait durablement réduire les croyances en des théories du complot. De même, avec le développement des IA conversationnelles, certaines personnes nouent de véritables liens affectifs avec ces systèmes, ce qui les rend vulnérables aux mises à jour de l'algorithme et potentiellement à des tentatives de manipulation.

Politique, travail, publicité, famille… la manipulation est partout. Nous distinguons la manipulation individuelle, qui cible une personne précise et tient compte de ses caractéristiques, de la manipulation des masses, présente notamment dans la publicité. Lorsqu'elle s'appuie sur des leviers sociologiques et psychologiques, celle-ci peut manipuler les comportements à grande échelle. Par exemple, Edward Bernays, célèbre publicitaire XX^{e} siècle, en mobilisant des médecins (figures d'autorité dans le domaine) pour promouvoir la consommation de bacon au petit-déjeuner, a modifié avec succès et durablement les habitudes alimentaires de la population des États-Unis (Bernays, 1928). Un exemple historique de manipulation à grande échelle est la propagande nazie orchestrée par Goebbels, qui utilisait des slogans simples et très répétés, contrôlait les moyens d'information et de divertissement, et disséminait des rumeurs pour influencer l'opinion publique. L'endoctrinement sectaire est un autre exemple de manipulation à grande échelle, les sectes utilisant des techniques d'influence et de contrôle de l'information pour inculquer leurs valeurs et croyances à leurs membres (Dubois, 1997).

Les diverses formes de manipulation et leurs outils :

Caire et Conchon (2018) répertorient diverses formes d'influence de l'attitude ou du comportement dans trois catégories : l'incitation (argumentation, suggestion, dissuasion, cooptation, corruption), la manipulation (déception, persuasion, désinformation) et la coercition (déterrence, intimidation, endoctrinement, subversion).

Joule et Beauvois (1987) citent de nombreuses méthodes de manipulation individuelle. Nous pouvons prendre les exemples suivants :

- Le pied dans la porte : un premier consentement de *« faible coût »*, facile à donner, est obtenu. Des conditions plus difficiles à accepter sont ensuite ajoutées. Le fait d'avoir obtenu un premier accord augmente la probabilité du second, qui aurait probablement été refusé si toutes les conditions avaient été annoncées dès le départ. Cette technique a été démontrée par Freedman & Fraser en 1966.
- L'amalgame cognitif : une personne ou un produit est associé à une image positive ou négative ; par exemple, une comparaison avec Hitler ou un rapprochement à des images de guerre et de fascisme. Même si la personne qui est confrontée à cette comparaison sait qu'elle est fausse, elle pourra tout de même être inconsciemment influencée.
- Soumission librement consentie (décrite par Joule & Beauvois, 2010) : une tâche paraît plus intéressante lorsque la personne a eu le sentiment (même illusoire) de l'avoir choisie librement. Ce phénomène repose sur le principe selon lequel les individus assument les actes qu'ils n'ont pas pu refuser.

Parfois, la manipulation peut simplement passer par la communication non verbale ou par les mots utilisés : des mots puissants inattaquables (comme *« liberté »*, *« égalité »* etc.), des euphémismes (*« conflit »* au lieu de *« guerre »*), des mots *« colorés »* qui évoquent l'expérience sensorielle (*« une vue à couper le souffle »*, *« de la nourriture qui met l'eau à la bouche »*) (Gass & Seiter, 2022). Ainsi, Walker et al. (2021) ont montré que les euphémismes ou dysphémismes peuvent affecter les jugements moraux.

Les techniques qui stimulent les émotions peuvent aussi avoir un effet important : les émotions étant nécessaires pour une prise de décision efficace, *« pour compromettre la capacité de prise de décision d'un adversaire, une approche est de manipuler et compromettre les émotions intervenant dans la prise de décision »* (Waltzmann, 2022 ; Pfister & Böhm, 2008).

La manipulation peut exploiter des failles cognitives et des leviers psychologiques. Par exemple, l'effet de primauté de l'information, où lorsque deux arguments contraires sont présentés dans un faible intervalle de temps, le premier tend à être perçu comme plus crédible. Rey et al. (2020) l'ont démontré dans une étude où des participants ont exprimé une nette préférence pour les voitures dont ils présentaient d'abord les informations positives, même lorsque l'ensemble des informations était identique. En revanche, dans d'autres conditions, l'effet de proximité peut être plus important : par exemple, une étude sur les annonces de dividendes et de résultats a montré que les informations communiquées en dernier ont un impact plus fort sur les rendements boursiers (Hartono, 2012). Un autre exemple est qu'une personne ciblée par une injonction a plus de chances d'y obéir si elle se sent libre de le faire. De même, la probabilité de répondre favorablement à une sollicitation augmente lorsque celle-ci provient d'un individu présentant des similarités avec la cible. Par exemple, dans une étude, un sondage par mail a reçu sept fois plus de réponses lorsque le mail était envoyé avec un nom et prénom d'expéditeur identiques à ceux du destinataire (Oates & Wilson, 2002). La distraction, l'isolement, la frustration, le mimétisme et d'autres leviers et failles cognitives peuvent être utilisés pour manipuler (Lorenc et al., 2021 ; Cacioppo & Hawkley, 2009 ; Fang et al., 2020 ; Hsu et al., 2018). L'utilisation de nudge, par exemple en rendant l'action souhaitée plus facile (comme en coupant les pommes à la cantine au lieu de les distribuer entières) ou plus séduisante, peut permettre de supprimer les obstacles qui empêchent l'action (Gass & Seiter, 2022). Il ne faut pas oublier de prendre en compte la culture de la cible, les spécificités du medium utilisé etc. (Géré, 2005).

Cependant, comme le rappelle Géré (2005), les techniques de persuasion et de manipulation ne sont pas automatiquement transposables d'un domaine à l'autre, les personnes traitant l'information différemment lorsque l'enjeu est important : par exemple, influencer une décision militaire est plus difficile qu'influencer une décision d'achat d'une marque plutôt qu'une autre.

De même, les résultats des campagnes d'influence peuvent être imprévisibles. Par exemple, en 1988, Hansen et al. ont mis en œuvre dans des écoles de Los Angeles un programme de prévention contre la drogue selon trois modalités : un programme social, un programme émotionnel, et aucun programme. Les résultats ont montré que les élèves ayant suivi le programme émotionnel ont fini par consommer davantage de cigarettes, d'alcool et de drogues que ceux n'ayant suivi aucun programme. L'intervention des chercheurs a donc conduit à une augmentation de la consommation de substances chez une partie des élèves, produisant un effet inverse à celui escompté.

Face à la diversité des techniques de manipulation et d'influence qui peuvent compromettre la qualité des décisions, il est essentiel de s'interroger sur les moyens de renforcer la résistance cognitive des individus et des organisations. Comment préserver la capacité à décider de manière lucide et efficace dans un environnement saturé d'informations et de tentatives d'influence ?

➢ Bilan du paragraphe 2.2 : les points-clés

La décision humaine est limitée par des contraintes cognitives : mémoire, attention, émotions, surcharge d'information, etc.

Les biais cognitifs sont des raccourcis mentaux inconscients qui peuvent fausser la perception, le raisonnement ou le jugement. Ces biais sont fréquents, prévisibles et exploitables : ils représentent une vulnérabilité dans les contextes d'influence ou de manipulation.

Les biais de la décision apparaissent à différentes étapes du processus décisionnel : collecte d'information, évaluation des options, passage à l'action. La manipulation peut agir à différents niveaux (message, émotions, contexte, structure de choix) en exploitant ces failles pour orienter ou inhiber une décision.

2.3 Comment protéger et améliorer la décision ?

Cette section présente plusieurs pistes pour protéger la prise de décision évoquées dans la littérature, telles que l'entraînement des décideurs, la formalisation de procédures, la réduction des biais cognitifs ou encore l'introduction de systèmes d'aide à la décision.

2.3.1 Entraînement et procédures

Charzat (2024) mentionne l'entraînement du décideur pour renforcer son intuition et protéger la décision par la connaissance de soi et de l'adversaire. Il souligne l'importance d'avoir conscience de l'existence de la guerre cognitive menée par l'adversaire. D'après lui, il faut favoriser *« la culture générale pour enrichir l'intuition de la plus grande variété de schémas et, plus largement, de l'expérience offerte par l'Histoire de l'humanité »*. Il mentionne également la délégation des décisions moins importantes, qui permet de

libérer l'attention et le temps du décideur qui pourra alors mieux prendre les décisions les plus importantes.

Nous avons cité dans le paragraphe *1.4 Représentations de la guerre cognitive* différentes matrices qui peuvent aider à reconnaître une attaque et à détecter les biais exploités par les adversaires, comme le framework DIMA. Ces outils sont très utiles pour contribuer à reconnaître une attaque et leur connaissance est importante pour les analystes qui cherchent à les détecter, comme pour les cibles potentielles qui peuvent s'appuyer dessus pour se sensibiliser à leurs propres biais cognitifs et pour aiguiser leur esprit critique. Ce dernier est une compétence primordiale pour protéger et améliorer la prise de décision dont les décideurs ne peuvent se passer. Il ne faut donc pas négliger le rôle de l'éducation et de la formation. Celle-ci doit être à la fois préalable à la décision et continue via des débriefings et retours d'expérience fréquents, l'analyse post-décision améliorant l'apprentissage et la qualité des décisions futures (Ellis et al., 2006). Dans certains contextes, il est même possible de simuler certaines situations afin d'améliorer la formation à la prise de décision et le retour d'expérience sans risques réels.

Nous avons également montré dans le paragraphe *1.6.8 Stratégies transversales* concernant les éléments de défense contre la guerre cognitive, que la mise en place d'entraînements cognitifs et de procédures qui permettent de vérifier la décision sont des étapes importantes pour rendre celle-ci plus résistante et résiliente face aux attaques cognitives et informationnelles. En effet, utiliser des procédures formalisées permet de compenser les effets du stress ou de la surcharge cognitive par exemple, réduisant les biais possibles ainsi que les oublis (Orasanu et al., 1993).

Une autre piste est de faire prendre la décision par un collectif plutôt qu'un individu isolé, car la diversité des points de vue peut améliorer la qualité du jugement (Surowiecki, 2004). Toutefois, les groupes augmentent le temps de prise de décision et peuvent introduire des biais tels que la polarisation ou la dilution de la responsabilité, qui peuvent réduire l'esprit critique et encourager la passivité (Sunstein, 1999 ; Darley & Latane, 1968). Pour limiter ces effets, il est essentiel de composer un groupe décisionnaire avec une diversité d'opinion, d'instaurer des mécanismes comme l'avocat du diable, des délibérations structurées ou le vote anonyme (Surowiecki, 2004 ; Nemeth, 1995). Ces méthodes favorisent la confrontation d'idées et la responsabilisation individuelle, renforçant ainsi la qualité et la fiabilité des décisions collectives.

2.3.2 Réduction des biais (débiaisage)

Le *« débiaisage »* (*debiasing* en anglais) ou réduction des biais, désigne l'ensemble des méthodes visant à atténuer ou corriger les erreurs systématiques de jugement dues aux biais cognitifs afin d'améliorer la qualité de la prise de décision. Elles sont d'autant plus importantes que les individus sont souvent inconscients de leurs propres biais et ont des difficultés à les corriger d'eux-mêmes : donner aux participants des informations sur leurs propres biais n'a pas d'impact sur leur scepticisme (List et al., 2024 ; Larrick, 2004).

Trois approches principales de débiaisage sont utilisées (Morewedge et al., 2015) : modifier les incitations, optimiser l'architecture du choix, former les individus à mieux décider. Parmi ces outils, il est important de noter que certains s'apparentent à de la manipulation. Il faut privilégier les approches qui aident le décideur à prendre conscience de ses biais, afin qu'il conserve sa liberté de choix (Lebraty, 2004).

Modifier les incitations :

Cette approche consiste à ajuster les récompenses ou les conséquences des choix pour motiver des comportements ou des décisions plus rationnels. Par exemple, réduire le prix des fruits frais dans une

cantine augmente leurs ventes. Cette méthode ne fonctionne pas pour tous les biais et soulève la problématique de la légitimité de ceux qui définissent les solutions optimales.

Optimiser l'architecture du choix :

Il s'agit de structurer la présentation des options et la manière dont les choix sont sollicités pour faciliter les décisions sans restreindre la liberté : c'est le nudging. Nous pouvons par exemple simplifier les informations disponibles ou mettre en valeur celles qui sont les plus importantes, proposer une option par défaut plus adaptée, choisir une échelle, une unité de mesure ou un référentiel plus adapté à sa compréhension, etc. Ces méthodes sont peu coûteuses et préservent la liberté de choix, mais ne suffisent pas toujours et ne s'appliquent pas à tous les biais.

Former à mieux décider :

Le débiaisage peut se faire par la formation à la pensée critique et la sensibilisation aux biais. Nous pouvons mentionner l'adoption de règles inférentielles telles que des perspectives alternatives (*« consider-the-opposite »* ou *« consider-an-alternative »*), le raisonnement statistique, la simulation d'analyses contradictoires, ou l'encadrement algorithmique de certaines décisions (Milkman et al., 2009 ; Soll et al., 2015).

L'entraînement peut aussi être efficace : par exemple, Franiatte et al. (2023) ont montré que la répétition de courts entraînements contre certains biais permet d'améliorer le raisonnement. Les jeux éducatifs interactifs ou les vidéos pédagogiques améliorent l'esprit critique et rendent plus prudent envers la désinformation potentielle, avec des effets durables et généralisables, tel que démontré par plusieurs expérimentations (List et al. 2024 ; Morewedge et al., 2015 ; Baker, 2010).

Approches combinées :

Enfin, des approches combinées, incluant l'usage d'outils et de procédures structurant la décision, renforcent l'identification et la correction des biais (Arkes, 1991 ; Larrick, 2004). Il est possible de mettre en place des listes de contrôle (par exemple, la checklist de détection des heuristiques et pièges décisionnels établie par Kahneman et al. en 2011) ou encore d'être attentifs à la diversité cognitive des équipes (Cristofaro, 2017). Nous pouvons également considérer l'intégration de modèles ou d'outils de contrôle de la qualité dans un Système d'Aide à la Décision (SAD) (Lebraty, 2004 ; Cristofaro, 2017).

2.3.3 Les systèmes d'aide à la décision

Un des moyens d'améliorer les processus décisionnels est d'implémenter des systèmes d'aide à la décision (SAD), ou des assistants intelligents qui permettent de réduire les biais et d'améliorer différentes étapes du processus décisionnel : collecte d'informations, tri des informations utiles, construction de la conscience de situation, recherche de solutions, choix de la solution optimale (Keen & Scott Morton, 1978)…

2.3.3.1 Les différents types de systèmes d'aide à la décision

Il existe des types de SAD très variés. Ils peuvent être implémentés dans divers types d'applications (mobile, web…), utiliser différentes méthodes de calcul (utilisation de poids, de critères de comparaison,

de règles de priorité, de règles de Bayes, etc.) et être appliqués dans des domaines variés (Umar et al., 2017). Ils peuvent être passifs (collection et organisation des données), actifs (traitement des données et proposition de solution automatisée) ou coopératifs (analyse et solution générées par le système, mais révisées par l'humain) (García-Alcaraz et al., 2023). Ils peuvent être basés sur des modèles analytiques, mathématiques ou statistiques, sur les données ou les connaissances, orientés sur la communication… Ils sont de plus en plus nombreux à incorporer de l'intelligence artificielle (machine learning, fouille de données etc.) à mesure que ce champ se développe (Onwujekwe & Weistroffer, 2025).

2.3.3.2 Objectif et conception d'un système d'aide à la décision

Un système décisionnel bien conçu peut gérer en autonomie de nombreuses étapes de la décision. Il est généralement utilisé comme assistant et fonctionne en collaboration avec l'humain. Ce dernier prend la décision finale mais peut aussi interagir avec le système tout au long du processus (en lui demandant des informations et en lui en fournissant). Ainsi, d'après Sedogbo et al. (2007), les SAD efficaces sont ceux qui rendent le problème transparent pour l'utilisateur. Lebraty (2004) ajoute qu'un SAD doit *« amener le décideur à être en adéquation avec son contexte ou sa tâche »*. Pour contribuer à la réduction des erreurs cognitives dans la décision, le système doit *« donner conscience au décideur que certains des raccourcis de raisonnement qu'il utilise l'ont conduit à dévier de ses intentions initiales »*.

Pour construire un système d'aide à la décision efficace, l'une des difficultés est de construire une conscience de situation fusionnant les différentes sources d'information disponibles, qui comprenne non seulement les objets, mais aussi les relations entre eux. Une autre difficulté est l'inclusion de différents types de connaissances : déclaratives (quoi faire, quelles sont les préférences), procédurales (comment le faire, quelles sont les contraintes de ressources) et opérationnelles (quand le faire) (Sedogbo et al., 2007).

Lebraty (2004) distingue deux grands types de systèmes d'aide à la décision, présentant chacun des atouts et des limites. Les premiers sont basés sur des modèles : algorithmes, règles de décision, logiques de raisonnement ou simulations. Leur principal avantage réside dans leur cohérence interne et leur explicabilité : ils reposent sur des principes connus et permettent de justifier les recommandations émises. En revanche, ils sont souvent rigides et peinent à s'adapter à des contextes nouveaux ou à des données incomplètes. Les seconds sont basés sur la manipulation de données, c'est-à-dire sur l'exploration et l'analyse automatique de grands volumes d'informations (fouille de données, apprentissage statistique, machine learning). Ces systèmes sont plus adaptatifs et capables de détecter des régularités inattendues, mais leur fonctionnement repose sur des corrélations plutôt que sur des relations causales explicites. Ils peuvent ainsi produire des résultats pertinents sans pouvoir en expliquer le raisonnement sous-jacent, ce qui limite la confiance que l'utilisateur peut leur accorder.

Enfin, Lebraty distingue deux types d'utilisateurs de ces systèmes : l'expert, qui recherche généralement un outil d'assistance capable d'enrichir son raisonnement, et le novice, qui attend un soutien plus prescriptif. Cette distinction peut être rapprochée des niveaux de contrôle décrits par Rasmussen et Hollnagel (cf. paragraphe *2.1.5 Les niveaux de contrôle*), qui influencent la manière dont l'utilisateur interagit avec le système, depuis un contrôle opportuniste fondé sur des heuristiques jusqu'à un contrôle stratégique fondé sur une compréhension complète de la situation.

2.3.3.3 Risques et biais liés aux systèmes d'aide à la décision

Ces systèmes d'aide, s'ils peuvent faciliter la décision et améliorer la solution adoptée comme le temps nécessaire à la prise de décision, peuvent cependant introduire de nouveaux risques ou biais (Baudel et al., 2021) :

- Biais d'automatisation, aussi nommé *« biais de complaisance »* ou *« biais d'autorité »* : il se manifeste lorsque le décideur accorde une confiance excessive au système automatisé et suit ses recommandations sans exercer de jugement critique. Ce phénomène résulte souvent d'une mauvaise calibration de la confiance entre la fiabilité réelle du système et celle perçue par l'utilisateur (Lee & See, 2004 ; Donnot et al., 2022). Lorsque la confiance augmente, le décideur peut réduire le champ de son savoir-faire utilisé (Amalberti, 1996 ; Lebraty, 2004), jusqu'à se retrouver *« hors de la boucle »* (*out-of-the-loop*), phénomène démontré par Endsley et Kiris (1995) dans une tâche de navigation automatisée. Le système doit assister l'humain, mais celui-ci doit conserver une implication dans la tâche et une conscience de la situation suffisantes pour être capable de reprendre la main et de prendre des décisions en connaissance de cause, dont il portera la responsabilité.
- Aversion à l'algorithme : à l'inverse, une confiance insuffisante conduit l'utilisateur à se méfier du système ou à ignorer ses recommandations, même lorsqu'elles sont plus fiables que son propre jugement. Cette sous-confiance (*undertrust*), décrite par Donnot et al. (2022), constitue le pendant du biais d'automatisation et illustre la nécessité d'un calibrage dynamique de la confiance.
- Fatigue décisionnelle : elle survient lorsque l'utilisation des aides à la décision s'accompagne d'attentes accrues de productivité, ce qui accroît la charge cognitive et réduit la vigilance du décideur.
- D'autres biais inhérents à la décision humaine interférer dans la collaboration avec un système d'aide : biais de conformité (si d'autres utilisent l'algorithme, je peux avoir confiance), biais d'attraction (notamment dans les interfaces visuelles), ou encore effet d'ordre, influençant la perception et la pondération des informations.

2.3.3.4 Les systèmes d'aide à la décision dans le milieu militaire

Les SAD avec intelligence artificielle sont déjà utilisés dans le domaine militaire. Nous pouvons donner l'exemple du système d'aide à la prise de conscience de situation ukrainien Delta, qui a été testé sur le terrain à l'occasion de la lutte contre l'offensive russe et dévoilé publiquement en 2022 (Danylov, 2022). Delta fournit une carte numérique pour le suivi des unités, une messagerie sécurisée, diffuse des vidéos en temps réel (drones par exemple) et aide à la planification d'opérations. Il s'appuie sur des sources civiles (eEnemy, Telegam bots) et alliées (imagerie satellite) pour enrichir la connaissance tactique. Cet outil contribue à une accélération de la prise de décision grâce à l'agrégation de données en temps réel, facilite le partage d'informations sur la situation tactique et aide à coordonner les forces sur le terrain (Soesanto, 2024).

Un autre exemple est le Maven Smart System (MSS NATO), développé par Palantir Technologies. Cet outil s'appuie sur l'intelligence artificielle pour accélérer et améliorer la décision des commandants et des opérateurs dans un contexte opérationnel, en intégrant et analysant des données provenant de sources variées (classifiées, ouvertes, structurées comme non structurées) (Jaesa, 2025).

Enfin, le système ANTICIPE, développé par THALES, utilise un arbre de détection d'informations critiques à trois niveaux et un système de fouille de données dans des sources variées pour détecter les signaux faibles et établir une meilleure conscience de situation et agir en support de la planification et la prise de décision militaire[3]. Ce système sera plus amplement étudié et utilisé dans l'expérimentation 2 (*Chapitre 4*).

2.3.3.5 Vers un système d'aide à la décision pour la guerre cognitive ?

Il est possible de s'inspirer des SAD militaires existants pour imaginer un système aidant à la décision ou à la surveillance des actions de guerre cognitive.

Un tel système devrait avoir accès aux données des réseaux sociaux, messageries sociales, médias de masse (radio, télévision, journaux…), événements géolocalisés et flux de données cyber. Il devrait utiliser l'apprentissage automatique et l'IA afin d'identifier les connexions et les schémas répétitifs et ainsi surveiller les campagnes de guerre cognitive. Il pourrait inclure un affichage de l'évolution des campagnes suspectes sous forme de cartes des lieux géographiques et virtuels (Cocron & Aronhime, 2022 ; Cao et al., 2021).

➢ Bilan du paragraphe 2.3 : les points-clés

La formation, l'entraînement et les procédures peuvent renforcer la qualité des décisions en réduisant les erreurs liées au stress, au temps, à la complexité ou aux attaques cognitives.

Le débiaisage repose sur plusieurs approches : modification des incitations, architecture des choix (nudging), sensibilisation et outils de raisonnement critique.

Les systèmes d'aide à la décision (SAD) peuvent accompagner l'humain dans la collecte, l'analyse et la hiérarchisation de l'information. Ces systèmes doivent être conçus pour soutenir et non remplacer le raisonnement humain : leur bon usage suppose un équilibre entre confiance et esprit critique. Une collaboration humain-machine efficace repose donc sur la transparence, la compréhensibilité et la capacité à repérer les biais ou limites du système.

2.4 Conclusion

Dans les deux premiers chapitres, nous avons défini la guerre cognitive, ses mécanismes, ses cibles et ses finalités, ainsi que les vulnérabilités de la prise de décision humaine qu'elle exploite. Dans un monde hyperconnecté où l'information circule sans filtre, cette nouvelle forme de guerre, insidieuse et transversale, bouleverse les repères traditionnels des conflits en agissant directement sur la cognition humaine. Elle agit en amont de l'action, sur les représentations mentales et les processus décisionnels, pour orienter, neutraliser ou manipuler les comportements d'individus ou de groupes, souvent à leur insu.

[3] https://www.youtube.com/watch?v=A2ZAHrT3UwM

La compréhension des processus décisionnels est donc essentielle pour anticiper et contrer les effets de la guerre cognitive. Le deuxième chapitre a montré que la décision humaine est soumise aux émotions du décideur, à des fragilités endogènes (biais et failles cognitives) et exogènes (influence et manipulation de la décision). Ces failles, inhérentes au fonctionnement du cerveau humain, peuvent être instrumentalisées dans des stratégies offensives. Par exemple, un agresseur pourra utiliser des informations cadrées de manière à déclencher des raccourcis de pensée, détourner l'attention, susciter des réactions émotionnelles ou créer une surcharge cognitive. La prise de décision repose également sur une construction dynamique de la conscience de la situation, qui peut elle-même être altérée par la qualité et l'accessibilité des informations disponibles.

Nous avons vu que les approches actuelles en matière de protection et support de la décision insistent sur la nécessité de l'éducation à l'esprit critique ou rationnel, de l'entraînement à la résilience cognitive et du recours à des systèmes d'aide à la décision. Ces derniers peuvent jouer un rôle crucial pour détecter les biais, structurer l'analyse, renforcer la conscience de la situation et soutenir le raisonnement dans des environnements incertains ou manipulés. Ils ont toutefois des limites et peuvent introduire leurs propres biais ou entraîner une dépendance excessive.

Étant donné la menace que représente la guerre cognitive, il est primordial de mieux la comprendre et la prendre en compte. Nous soulignons l'importance de développer des outils aptes à repérer les influences cognitives à l'œuvre et de renforcer la robustesse des processus décisionnels, en particulier chez les décideurs civils et militaires. Les éléments théoriques exposés ici nourrissent une réflexion sur la conception de dispositifs technologiques permettant la défense ou l'offensive dans le champ cognitif.

PARTIE II – Études empiriques

Comment concevoir un système d'aide à la décision capable de détecter, caractériser et contrer les actions de guerre cognitive qui visent les décideurs civils et militaires ?

L'objectif de ces travaux de doctorat est d'établir les fondements préliminaires à la conception d'un système d'aide à la décision, qui collecte et analyse les informations nécessaires à une prise de décision défensive ou offensive dans le contexte de la guerre cognitive.

Dans ce but, nous avons déjà examiné dans la première partie (*Chapitre 1* et *Chapitre 2*) de ce manuscrit (Figure 1 – p. 16) :

- **Revue de littérature sur la guerre cognitive :** Qu'est-ce que la guerre cognitive ? Comment la définir ? Quels sont ses outils ? Qui sont les acteurs ? Quels sont les enjeux ?
- **Revue de littérature sur la prise de décision :** Qu'est-ce que la décision ? Qu'est-ce qui peut la mettre en danger ? Comment la protéger ?

Nous étudierons dans la suite de ce manuscrit (Figure 1) :

- **Quels outils peuvent être utilisés pour influencer la décision ?** Dans le *Chapitre 3*, nous prenons l'exemple des éléments qui rendent un texte plus crédible et influencent donc la perception de l'information et la prise de décision qui en découle.
- **Comment mettre en place un outil d'aide à la décision dans le contexte de la guerre cognitive ?** Dans le *Chapitre 4*, nous étudions la mise en place d'un outil d'aide à la décision qui permette d'améliorer la conscience de situation et de proposer des solutions pour influencer la cible, avec l'exemple d'un jeu de stratégie semi-collaboratif.
- **Qui sont les cibles privilégiées des actions de guerre cognitive ?** Dans le *Chapitre 5*, nous nous penchons sur la qualification des individus qui les rendent pertinents ou faciles à cibler via des actions de guerre cognitive.

Chapitre 3 - Facteurs d'influence de la décision - Évaluation de la crédibilité perçue de messages textuels dans le contexte de la désinformation

Chacun de nous prend des décisions quotidiennement, au terme d'un processus plus ou moins conscient de prise d'information et d'évaluation des options disponibles. La sphère numérique nous apporte une masse d'informations toujours plus conséquente et l'intelligence artificielle générative facilite leur manipulation et diffusion. Dans ce contexte, il est important de comprendre ce qui rend ces informations crédibles aux yeux de leurs destinataires et ce qui en fait un levier d'influence, notamment pour la guerre cognitive.

Cette étude explore la perception de la crédibilité des informations reçues sous forme de textes, qu'elles soient vraies ou manipulées. Nous avons identifié un grand nombre de facteurs de crédibilité grâce à une revue de littérature. Par la suite, des réunions d'experts et la méthode du tri de cartes nous ont permis d'isoler les facteurs individuellement les plus influents.

Les résultats de cette expérimentation montrent que les facteurs tels que la présence de détails, d'exemples, de sources et de répétitions augmentent la crédibilité d'un texte. Tandis que les émotions, l'assertivité et l'intention manifeste de convaincre y contribuent moins. De même, les participants avec un niveau d'études élevé ou une faible sensibilité à la désinformation confèrent plus de crédibilité à un texte citant des sources. Ces résultats soulignent l'importance de former les citoyens à l'esprit critique pour mieux résister à la désinformation dans un environnement saturé d'informations manipulées.

3.1 Introduction

Dans le contexte contemporain, la guerre cognitive émerge comme une menace significative, utilisant divers moyens pour déstabiliser l'adversaire en altérant sa cognition et ses représentations mentales (Bernal et al., 2020 ; Claverie et al., 2021). L'information et sa manipulation sont un des outils de la guerre cognitive, et la décision en est souvent une cible.

Or, la prise de conscience de la situation est la première étape de la prise de décision, incluant la prise d'information et son interprétation (Huber et al., 2011 ; Endsley & Garland, 2000). Une conscience de la situation juste et précise est essentielle pour prendre des décisions adaptées. Cette étude porte sur l'interprétation des informations disponibles qui interviennent dans la chaîne de décision et contribuent à la construction de la conscience de situation.

Nous cherchons à identifier les critères influençant la compréhension de la situation à travers la lecture d'un texte, ce qui permet à une personne de construire une représentation de la situation en fonction de ses connaissances et de sa perception. Nous nous intéressons à la manière dont une personne évalue une information vraie ou manipulée, et à l'impact de cette évaluation sur une éventuelle décision. Nous cherchons également à déterminer si certaines personnes accordent plus de confiance à certains facteurs, en fonction : de leur profil démographique, de leur niveau d'études ou encore de leur niveau de sensibilité à la désinformation.

3.2 Cadre théorique : les facteurs de crédibilité d'un texte d'après la littérature

Nous supposons que la crédibilité perçue d'un texte est un facteur important déterminant son potentiel d'influence. En effet, Li et Suh (2015) remarquent que la crédibilité de l'information est un prédicteur important de l'action ultérieure du consommateur de l'information.

Nous avons mené une revue de la littérature anglaise et française portant sur la crédibilité perçue de textes en fonction de leur contenu et de leur mode de présentation. Nous estimons que la littérature dans ces deux langues nous permet de couvrir le cadre culturel pertinent pour étudier la guerre cognitive sous un angle défensif.

Cette revue de littérature nous a permis de mettre en évidence 39 facteurs influençant cette crédibilité (Tableau 10) : des facteurs associés au format et au langage du texte tels que la présence d'émotions ou la ponctuation, des facteurs associés à l'auteur et aux relais du message comme leur expertise ou leur affiliation, ainsi que des facteurs associés au média et au lecteur lui-même (Baptista & Gradim, 2020 ; Dietrich, 2010 ; Hansen et al., 2011 ; Henderson & Nutt, 1980 ; Keshavarz, 2020 ; Metzger et al., 2003 ; Rastogi & Bansal, 2023 ; Shariff, 2020 ; Wathen & Burkell, 2002). Parmi ces nombreux facteurs, certains ont déjà été comparés les uns aux autres dans la littérature, ce qui permet d'isoler un sous-ensemble contenant les plus significatifs. Cependant, nous n'avons pas pu identifier de travaux antérieurs comparant entre eux les facteurs inhérents au texte ayant le plus d'influence sur la crédibilité perçue du texte par son lecteur.

L'intérêt d'établir un classement est multiple : il permet de distinguer les facteurs qui sont les plus importants à introduire dans un message écrit pour optimiser sa crédibilité, mais aussi lesquels doivent éveiller la méfiance du lecteur ; il peut contribuer à mieux comprendre la perception humaine de l'information écrite. Ce classement pourra également servir de base pour des recherches futures étudiant les interactions entre les facteurs apportant le plus de crédibilité à un message.

Suite à cette revue de littérature, nous avons classé les facteurs d'influence identifiés en cinq catégories principales :

- crédibilité du message ;
- crédibilité de l'auteur du message ;
- crédibilité du média ;
- caractéristiques du lecteur ;
- caractéristiques de l'environnement.

Ainsi, nous avons établi une première liste de 39 facteurs (Tableau 10), chacun comportant généralement de deux à cinq modalités. Ce nombre de facteurs est trop important pour notre expérimentation, il faudra le réduire par la suite (paragraphe *3.4.1 Phase A1 : Sélection des facteurs les plus pertinents*).

Tableau 10 : Liste des facteurs influençant la crédibilité perçue d'un texte d'après la littérature étudiée

Facteur d'influence	Modalités	Dimensions	Sources
Crédibilité du message (23 facteurs)			
Émotions (ou cadrage : perception de gain apporté ou perte évitée)	Positive / négative / neutre *(leur efficacité diverge selon les études)*	Émotions	Hansen et al., 2011 ; Wathen & Burkell, 2002
Contenu du message	Portée / profondeur	Qualité information	Wathen & Burkell, 2002
Erreurs de français et d'orthographe	Présence d'erreurs / absence d'erreurs	Soin de la présentation	Keshavarz, 2020
Ponctuation	Ponctuation neutre / "!" / "?" / "..."	Émotions	
Émoticônes	Émoticônes positifs / négatifs / absents		
Clarté, lisibilité du message	Clair / pas clair - le score *SMOG* permet d'évaluer la clarté d'un texte (Mc Laughlin, 1969)	Soin de la présentation	
Objectivité de l'information	Présence / absence d'intention de convaincre ou de vendre	Qualité information / émotions	
Qualité de langage	Langage familier / langage soutenu	Soin de la présentation	Shariff, 2020 ; Keshavarz, 2020
Répétitions	Présence / absence de répétitions		
Autres caractéristiques linguistiques	Écriture inclusive ou non / hashtags		
"Puissance" du langage utilisé	Langage puissant, assertif (plus crédible) / neutre / faible, hésitant (moins crédible)	Émotions	Metzger et al., 2003
Type d'information	Information vraie / désinformation / mésinformation	Qualité information	Rastogi & Bansal, 2023
Caractéristiques du message	"Prétentieux" / persuasif / informel / suggestif	Émotions	Baptista & Gradim, 2020
Orientation du message	Positif / négatif envers le sujet traité		Shariff, 2020
Date du message	Information récente / ancienne	Qualité information	
Sources	Présence / absence de sources		Keshavarz, 2020 ; Kondamudi et al., 2023
Chiffres	Présence / absence de chiffres ou statistiques		Keshavarz, 2020

Facteur d'influence	Modalités	Dimensions	Sources
Exemples	Présence / absence d'exemples	Qualité information	Slater & Rouner, 1996 ; Wathen & Burkell, 2002 ; Keshavarz, 2020
Détails	Présence / absence de détails		Wathen & Burkell, 2002 ; Keshavarz, 2020
Moyen de communication	Audio / vidéo / image / longueur du message	Présentation	Wathen & Burkell, 2002 ; Keshavarz, 2020 ; Kondamudi et al., 2023
Message sur les réseaux sociaux : nombre de réactions	Peu / beaucoup de réactions, commentaires & partages	Émotions / biais de conformité au groupe	Shariff, 2020 ; Keshavarz, 2020
Message sur les réseaux sociaux : qualité des réactions	Commentaires et réactions positifs / neutres / négatifs		
Familiarité du sujet avec le contenu du message	Information jamais rencontrée / déjà rencontrée / bien connue / en contradiction avec une information déjà connue	Émotions / biais de confirmation / biais de simple exposition	Metzger et al., 2003
Crédibilité de l'auteur du message (7 facteurs)			
Sexe, âge	Femme / homme, âge	Préjugés	Metzger et al., 2003 ; Shariff, 2020
Localisation géographique / affiliation de l'auteur	Entité neutre / entité alliée / entité adverse	Émotions / préjugés	Wathen & Burkell, 2002 ; Metzger et al., 2003 ; Shariff, 2020 ; Keshavarz, 2020
Expertise / éducation	Expert sur le sujet / haute éducation sans rapport / faible éducation	Qualité information / préjugés	
Autres caractéristiques de l'auteur	Ethnie, fiabilité…	Émotions / préjugés	
Caractéristiques de l'auteur sur les RS	Nombre de followers (RS) / d'amis / de posts / fréquence des posts / âge du compte	Biais de conformité au groupe	Shariff, 2020 ; Keshavarz, 2020
Rapport du lecteur à l'auteur du message	Sympathie du lecteur envers l'auteur / points communs perçus	Émotions / préjugés	Wathen & Burkell, 2002 ; Metzger et al., 2003 ; Kuutila et al., 2024
Attitude du lecteur envers le contenu	Contradiction / accord / neutre (sujet non connu)	Qualité information / préjugés	

Facteur d'influence	Modalités	Sources
Crédibilité du média (3 facteurs)		
Canal de transmission du message	Media traditionnels / réseaux sociaux / communication directe (messagerie privée) / site web	Wathen & Burkell, 2002
Dynamisme & interactivité du média		Metzger et al., 2003
Utilisabilité du média	Facilité d'utilisation / accessibilité / sécurité / engagement utilisateur	Keshavarz, 2020
Caractéristiques du destinataire/lecteur (3 facteurs)		
Préconceptions ou attentes	Dues par exemple aux consignes données au préalable ou à l'environnement	Endsley & Garland, 2000
Facteurs internes au lecteur	Stress / fatigue / charge cognitive / état émotionnel / maîtrise du sujet / personnalité, valeurs, expérience ou conviction individuelle / sensibilité à la désinformation / âge / catégorie socio-professionnelle / intérêt pour l'information, implication et motivation voire besoin de l'information / idées préconçues sur l'information ou la source…	Dietrich, 2010 ; Wathen & Burkell, 2002
Style cognitif du lecteur	Aversion ou tolérance au risque	Henderson & Nutt, 1980
Caractéristiques de l'environnement (3 facteurs)		
Facteurs externes	Présence de distractions, d'incertitude et d'hostilité dans l'environnement	Elbanna & Child, 2007
Temps	Temps passé depuis que l'information a été rencontrée	Wathen & Burkell, 2002
Support du message	Papier (magazine, journal, livre, tract…) / digital	

Parmi ces facteurs, le score SMOG est une formule de lisibilité estimant le niveau de lecture qu'une personne doit avoir atteint pour comprendre pleinement un texte donné. Il est obtenu à partir du nombre de mots de trois syllabes ou plus dans un texte (Mc Laughlin, 1969). Nous considérons que maximiser le nombre de mots composés de trois syllabes ou plus revient à maximiser le score SMOG du texte.

3.3 Question de recherche

Nous formulons l'hypothèse que certains facteurs, tels que la présence d'émotions ou le soin apporté à la rédaction (langage soutenu), donnent plus de crédibilité à un texte que d'autres facteurs et lui confèrent un potentiel d'influence plus important. Nous souhaitons également étudier les corrélations entre les différentes caractéristiques des participants à nos expérimentations, notamment en ce qui concerne les corrélations entre le niveau de sensibilité à la désinformation (score MIST – Maertens et al., 2023 – présenté dans le paragraphe *3.4.4 Phase B2 : Tri de cartes en ligne*) et les caractéristiques telles que l'âge

ou le plus haut diplôme obtenu. Au regard de l'hypothèse formulée et des résultats de notre étude de la littérature, nous tentons de répondre à plusieurs questions de recherche :

- Quels facteurs confèrent le plus de crédibilité à un message textuel ?
- Ces facteurs fonctionnent-ils pour un texte relevant de la vraie information comme pour un texte relevant de la désinformation ?
  - Nous cherchons ici à établir un classement des facteurs de crédibilité les plus importants afin de mieux identifier lesquels il faut chercher à distinguer dans des textes cherchant à désinformer.

- Y a-t-il des corrélations entre les caractéristiques des participants, notamment entre le niveau de sensibilité à la désinformation (score MIST) et des caractéristiques telles que l'âge ou le plus haut diplôme obtenu ?
  - La présence de telles corrélations pourrait donner des indications pour répondre à la dernière question de recherche ci-dessous.

- Y a-t-il une différence entre les classements qu'effectuent les participants en fonction de leur profil démographique, leur niveau d'études ou encore leur niveau de sensibilité à la désinformation ?
  - Nous cherchons ici à identifier si certains profils sont plus sensibles à différents éléments de désinformation. Cela pourrait nous permettre de distinguer certains profils démographiques plus influençables.

L'étude se déroule en trois étapes principales :

- La première est la sélection d'un petit ensemble de facteurs de crédibilité parmi les plus importants d'après la littérature, par des réunions d'experts. En effet, les 39 facteurs identifiés et leurs nombreuses modalités ne peuvent pas tous être testés.
- La deuxième étape consiste en la conception de textes illustrant ces facteurs et leur validation via un *tri de cartes*[4] (Rugg & McGeorge, 1997) au format papier.
- La troisième étape repose sur un tri de cartes numérique, durant lequel les participants doivent ordonner les textes du plus au moins crédible afin de nous aider à déterminer quels facteurs ont le plus grand potentiel d'influence.

3.4 Protocole expérimental

Le protocole mis en œuvre pour notre approche expérimentale se déroule en trois phases (Figure 8) :

- Phase A1 : sélection des facteurs à évaluer.
  1. Sélection des facteurs les plus influents dans la littérature.
  2. Réunions d'experts pour réduire le nombre de facteurs à étudier et isoler un petit nombre de facteurs expérimentaux.

[4] Tri de cartes : méthode qualitative où l'on demande à des participants de regrouper ou classer des items (représentés par des cartes) afin de révéler l'organisation ou la perception de ces éléments.

- Phase A2 : préparation du matériel pour la phase B et validation par un tri de cartes papier.
    1. Création de trois textes de base.
    2. Conception des cartes (cf. Figure 9) : trois séries de 12 textes illustrant chacun des 12 facteurs sélectionnés lors de la phase A1.
    3. Réunions d'experts pour évaluer et améliorer les textes.
    4. Tri de cartes au format papier (21 participants, 36 cartes) pour valider l'association adéquate entre chaque facteur et le texte qui l'illustre.
- Phase B1 : lancement du tri de cartes en ligne.
    1. Recrutement de 100 participants au minimum.
    2. Création de l'outil de tri de cartes en ligne.
    3. Envoi du tri de cartes aux participants.
- Phase B2 : tri de cartes en ligne.
    1. Consignes et explication du contexte, information RGPD (5 minutes).
    2. Questionnaire démographique (2 minutes).
    3. Trois tris de cartes : un pour chaque série de 12 textes.
    4. Questionnaire MIST (Maertens et al., 2023) : mesure de la sensibilité aux fausses informations (2 à 5 minutes).

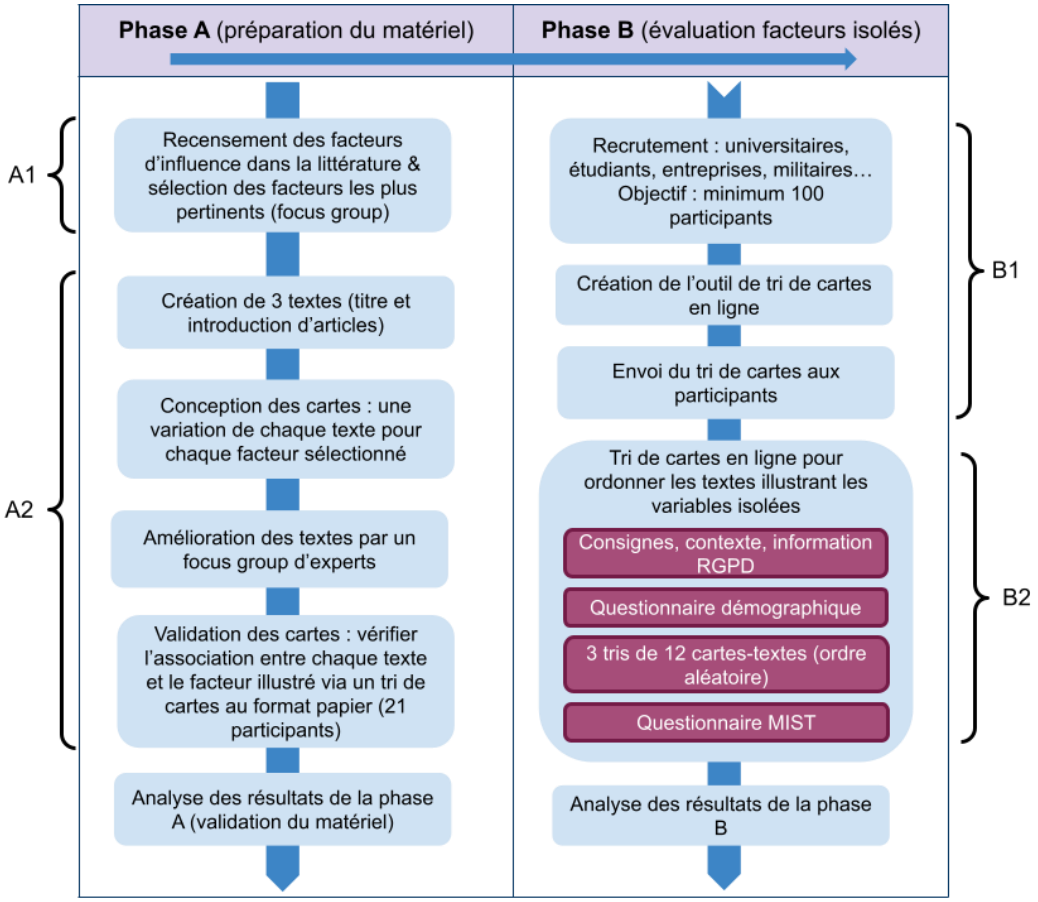

Figure 8 : Représentation schématique du protocole expérimental

3.4.1 Phase A1 : Sélection des facteurs les plus pertinents

Après avoir collecté les facteurs dans la littérature (voir paragraphe *3.2 Cadre théorique : les facteurs de crédibilité d'un texte d'après la littérature*), nous avons réduit la liste initiale de 39 facteurs (Tableau 10) à 12 facteurs expérimentaux (liste *Facteurs conservés* ci-dessous). Cette réduction a été réalisée en éliminant ceux qui étaient jugés par les experts comme moins pertinents ou trop complexes à tester.

Facteurs retirés :

- Médias visuels : la présence de vidéos ou d'images en lien avec le message a été retirée, car bien que des images puissent être générées automatiquement, leurs caractéristiques (couleurs, éléments représentés) pouvaient influencer la crédibilité de manière imprévisible.
- Facteurs contextuels : d'autres facteurs retirés incluent les critères MICE (Money, Ideology, Compromission, Ego), la présence de distractions environnementales (l'expérimentation étant en ligne), ainsi que l'utilisabilité et l'interactivité du média.
- Auteur et média : nous avons choisi de retirer les facteurs concernant l'auteur, le média utilisé et les caractéristiques du compte de l'auteur sur les réseaux sociaux, car ceux-ci introduisaient une variabilité et des effets d'interaction importants.

Stratégies de réduction des modalités :

Quatre stratégies ont été utilisées pour simplifier et réduire les modalités des facteurs restants :

1. Simplification des modalités : par exemple, au lieu de lister un large spectre d'émotions (colère, aversion, frustration, étonnement, admiration…), nous avons conservé uniquement les émotions positives et négatives.
2. Réduction des modalités : Nous avons testé uniquement les modalités associées à une meilleure crédibilité d'après la littérature, comme l'absence d'erreurs d'orthographe dans les textes plutôt que la présence d'erreurs.
3. Suppression de facteurs : les facteurs jugés moins pertinents pour un contexte de guerre cognitive, trop complexes à tester, ou encore les moins susceptibles d'influencer ont été supprimés. Par exemple, nous n'avons pas conservé la familiarité du lecteur avec le contenu du message.
4. Rapprochement de facteurs : certains facteurs proches ont été fusionnés, comme l'usage d'un langage assertif et de points d'exclamation.

Facteurs conservés :

Après réduction, les 12 facteurs de crédibilité d'un message textuel retenus pour l'expérimentation sont les suivants :

- Émotions positives
- Émotions négatives
- Langage soutenu
- Langage & ponctuation assertifs (!)
- Langage & ponctuation suggestifs (?)
- Présence de répétitions
- Message pas clair (score SMOG élevé)
- Présence apparente d'intention de convaincre
- Présence de nombres
- Présence d'exemples
- Présence de détails
- Présence de sources

Le facteur de type d'information (vraie information ou désinformation) a été réintroduit par la suite avec les trois différents textes évalués (voir paragraphe *3.4.2.1*).

3.4.2 Phase A2 : Préparation du matériel

Au cours de la phase A2, nous abordons la préparation du matériel expérimental en vue de la phase B de notre approche. Nous avons retenu la méthode du tri de cartes afin de répondre aux questions de recherche énoncées (Rugg & McGeorge, 1997). Ainsi, en fin de phase A2, nous demanderons à une vingtaine de participants d'associer des cartes « message » à des cartes « définition ». Les cartes « message » sont constituées d'un support papier imprimé comprenant des titres et des phrases introductives d'articles ; les cartes « définition » contiennent un intitulé de facteur et sa définition (cf. Figure 9 et Figure 10). En phase B, nous utiliserons une version numérique des cartes « message » que les participants devront ordonner par crédibilité croissante. Ces cartes présentent des variations qui illustrent les facteurs et modalités à évaluer.

Étant donné le nombre relativement élevé de facteurs retenus (12), nous avons choisi de tester chacun d'entre eux individuellement, tout en assumant que certains effets d'interaction entre facteurs pourraient ainsi être négligés. Chaque carte correspond donc à un facteur distinct.

3.4.2.1 Création de 3 textes : Sélection de titres et introductions d'articles

Pour évaluer l'influence du contexte sur l'ordonnancement des cartes, nous avons choisi plusieurs titres et introductions d'articles de type « journal ». Ainsi, trois scénarios de base (cartes comprenant titre et introduction d'article) ont été créés pour fournir suffisamment de matière tout en évitant d'allonger excessivement la durée et la complexité de l'expérimentation pour les participants. Le format court et accessible, centré sur l'actualité, correspond à un mode de consommation d'information rapide, par exemple via les réseaux sociaux (62% des français s'informent aujourd'hui via les réseaux sociaux – Patino, 2023).

L'objectif étant d'évaluer les facteurs d'influence les plus importants dans un contexte de guerre cognitive, nous avons fait une pré-sélection d'une vingtaine d'articles d'actualités rédigés en français, en lien avec ce sujet. Les textes choisis traitent de désinformation et d'influence, et certains présentent de fausses informations. Parmi ces textes, afin que les participants se sentent concernés par l'information et investis, nous en avons sélectionné 3 qui sont liés à la France, à des outils (médias) utilisés quotidiennement par de nombreux français ou à des célébrités, tout en traitant de sujets variés.

La sélection finale comprend trois articles (scénarios) incluant chacun un titre et une introduction. Deux articles sont de véritables titres et introductions (choisis dans la presse de langue française), le troisième étant traduit de l'anglais ; quelques modifications mineures ont été introduites pour utiliser un langage plus neutre. Le premier article décrit des faits réels et récents en France. Le deuxième article combine des faits réels avec des projections futures et des inquiétudes subjectives. Le troisième article présente des informations mensongères conçues pour tromper les lecteurs.

Les articles sont les suivants :

1. **Pochoirs d'étoiles de David à Paris : la piste d'une opération d'ingérence russe privilégiée.** Les actions auraient été commanditées par l'homme d'affaires moldave Anatolii Prizenko et largement relayées par le réseau de propagande prorusse Doppelgänger. Deux couples moldaves sont fortement suspectés d'être à l'origine d'une partie des 250 pochoirs. *(source :*

www.lemonde.fr/societe/article/2023/11/07/pochoirs-d-etoiles-de-david-a-paris-la-piste-d-une-operation-d-ingerence-russe-privilegiee_6198775_3224.html)

2. **Les algorithmes qui pourraient élire le prochain président.** Les campagnes d'Obama et de Trump et le piratage massif de courriels contre Macron montrent comment la science des données peut influencer les élections démocratiques. Par ailleurs, le Parti Synthétique au Danemark, dirigé par un chatbot nommé Leader Lars, soulève des questions liées à l'émergence de gouvernements pilotés par l'IA. *(source : https://english.elpais.com/science-tech/2023-01-20/the-algorithms-that-could-elect-the-next-president.html)*

3. **Taylor Swift et Selena Gomez dénoncent l'aide occidentale à l'Ukraine.** Taylor Swift affirme que les Ukrainiens se comportent de manière discutable concernant les fonds attribués par l'Occident. Quant à Selena Gomez, elle estime que lorsque les ukrainiens reçoivent de l'argent, leur situation ne s'améliore pas. *(source : https://korii.slate.fr/et-caetera/russie-utilise-fausses-citations-stars-propagande-anti-ukraine-doppelganger-desinformation-internet-reseaux-facebook-guerre?utm_source=pocket-newtab-fr-fr)*

3.4.2.2 Conception des cartes pour le tri de cartes

Une fois les scénarios de base (articles) sélectionnés (version « neutre »), des variantes ont été générées à l'aide de l'intelligence artificielle générative Chat-GPT, en lui fournissant les versions originelles des textes avec pour consigne « d'ajouter des émotions positives » ou autres facteurs, « sans modifier le sens du texte ». Elles ont ensuite été modifiées et améliorées manuellement. Elles implémentent les 12 facteurs présentés dans la section *3.4.1*. Ainsi, trois séries de 12 cartes représentant chacun des facteurs de crédibilité du texte (cartes « message ») ont été créées. Chaque série est basée sur l'un des trois scénarios « neutres » présentés dans la section *3.4.2.1*.

Les cartes ont été conçues pour ressembler à des notifications sur un écran de téléphone verrouillé, afin de plonger les participants dans un contexte d'information quotidien (voir Figure 9). Cette apparence n'a pas été conservée pour la phase B (tri de cartes en ligne) en raison de problèmes de lisibilité et d'affichage sur les écrans d'ordinateur.

Les cartes « message » sont accompagnées de cartes « définition » utilisées lors de la phase de validation du matériel (tri de cartes papier avec une vingtaine de participants pour valider les cartes « message »). Ces cartes comportent chaque facteur et sa définition.

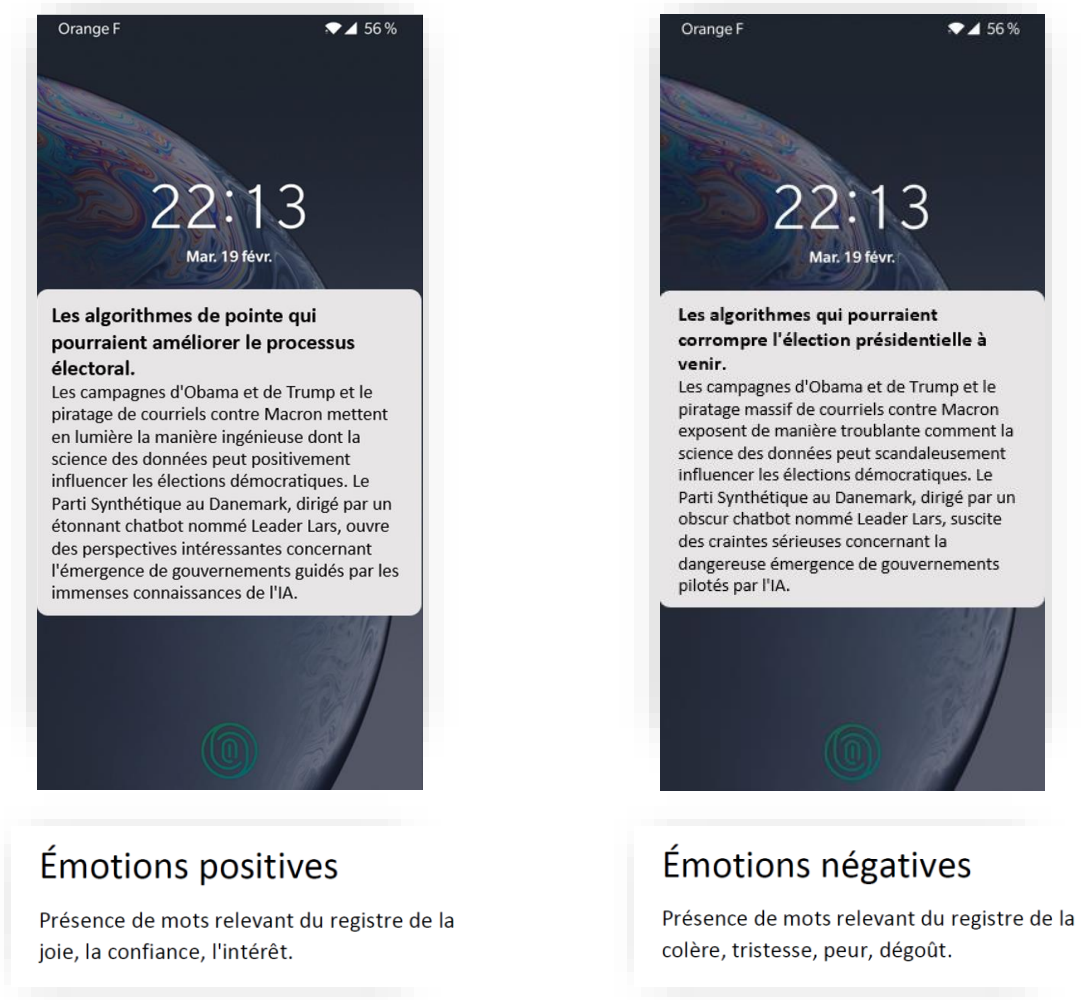

Figure 9 : Exemples de cartes implémentant : le facteur « émotions positives » pour le scénario 2 (en haut à gauche) et sa carte « définition » associée (en bas à gauche) ; et le facteur « émotions négatives » pour le scénario 2 (en haut à droite) et sa carte « définition » associée (en bas à droite)

3.4.2.3 Validation des textes par une réunion d'experts

Les cartes créées ont été validées par une réunion d'experts, afin de vérifier que chaque facteur était correctement illustré par le contenu du texte. Les experts ont également validé les cartes définissant les labels (soit les noms des facteurs) associés aux cartes « message ». Les définitions finalement adoptées pour les cartes « définitions » sont les suivantes :

Neutre : Émotions neutres, ponctuation neutre, langage clair et sans erreurs, pas de répétitions, absence ou faible présence des autres facteurs.

Émotions positives : Présence de mots relevant du registre de la joie, la confiance, l'intérêt.

Émotions négatives : Présence de mots relevant du registre de la colère, tristesse, peur, dégoût.

Message pas clair : Message au score SMOG élevé : phrases longues, présence de propositions subordonnées ainsi que de nombreux mots à 3 syllabes ou plus qui rendent le sens du message difficile à saisir.

Intention de convaincre : Présence d'une prise de parti apparente ; il peut y avoir un appel aux émotions négatives ou positives suivant l'orientation du parti pris,

Langage soutenu : Par opposition au langage familier ; présence de mots rares et recherchés sans pour autant altérer la clarté du message.

Langage & ponctuation assertifs : Langage puissant et assertif : par opposition au langage faible ; sans hésitations, présence d'affirmations fortes, qui ne laissent pas de place à l'interprétation.

Langage & ponctuation suggestifs / interrogatifs : Langage hésitant et suggestif : langage non assuré, présence de suggestions et éventuellement de questionnements.

Répétitions : Redondance de l'information et présence de répétitions de mots, d'adjectifs et d'expressions non neutres, qui appuient le parti pris dans le message / l'information que l'auteur souhaite faire passer.

la présence d'argumentation, ainsi que de mots relevant du registre de la négation du doute.

Nombres : Présence de données chiffrées et de statistiques précises qui soutiennent les allégations avancées dans le message.

Exemples : Présence d'exemples qui soutiennent les allégations avancées dans le message : événements, anecdotes particulières, objets, personnes, idées etc. qui représentent un extrait et non la totalité des éléments présentés.

Détails : Présence de détails donnant des informations supplémentaires sur la situation décrite.

Sources : La source de l'information est identifiée (qui l'a dit / qui l'a relevé / qui l'a relayé

Par la suite, deux pré-tests ont été conduits avec des experts neutres (non impliqués dans la création des cartes). Ces participants ont associé chaque carte à un label et fourni des retours sur la formulation, permettant des améliorations incrémentales.

Ces deux étapes ont abouti à une version stable des cartes « message ». Dans le paragraphe suivant, elles seront validées par une étape de tri de cartes au format papier.

3.4.2.4 Validation des cartes par un tri de cartes au format papier

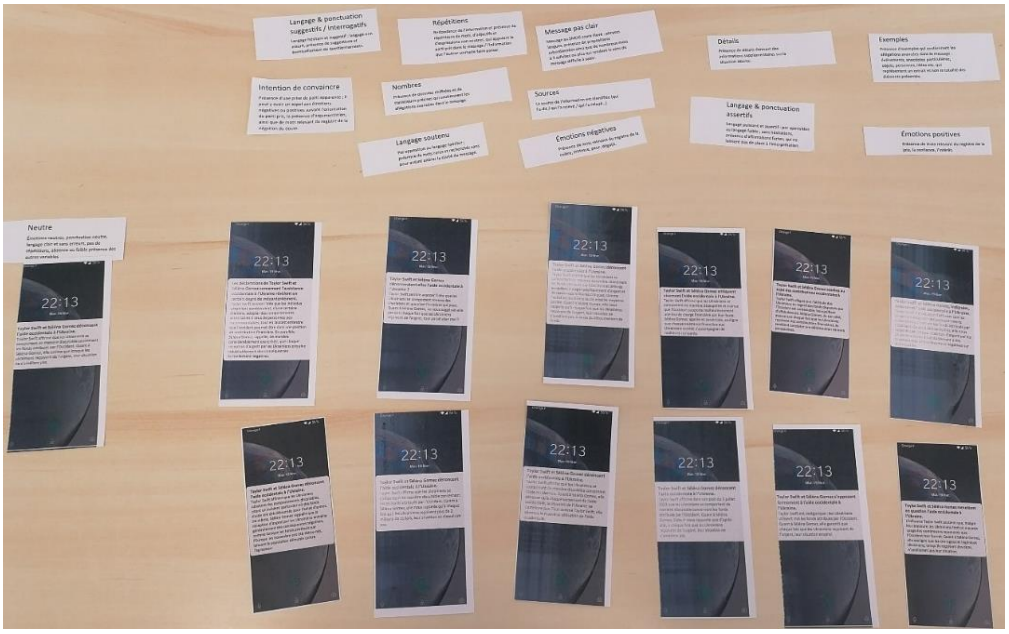

Figure 10 : Disposition initiale d'un tri de cartes (à gauche la carte neutre, à droite en haut les cartes définition, à droite en bas la série de 12 cartes illustrant les 12 facteurs possibles).

Un tri de cartes au format papier a été réalisé en présentiel avec 21 participants pour évaluer les cartes « message ». Sur la base des cartes neutres (3 cartes pour 3 scénarios), chaque participant a effectué trois tris, avec pour consigne d'associer chaque carte d'un groupe de 12 cartes « message » (en bas sur la Figure 10) à l'une des 12 cartes « définition » (en haut sur la Figure 10). La carte « neutre » était clairement identifiée dès le début de chaque tri afin de fournir un élément de référence pour comparer les autres cartes. L'objectif était donc de tester si les participants retrouvaient bien dans chaque texte conçu le facteur que nous avons cherché à illustrer dans le texte en question, avant d'utiliser ces cartes pour la phase B (tri de cartes en ligne).

Présentation des cartes :

Afin de s'assurer que chaque texte soit lu avant d'être trié, nous avons fait le choix de présenter aux participants chaque lot de 12 cartes un par un plutôt que simultanément les 3 lots de cartes associées à un facteur commun (voir Figure 11 pour la comparaison des deux options). Cette méthode a été choisie après un test avec les experts, qui a révélé que lorsqu'elles sont présentées par groupes de trois, les participants ne lisaient souvent que la première carte avant de catégoriser le groupe.

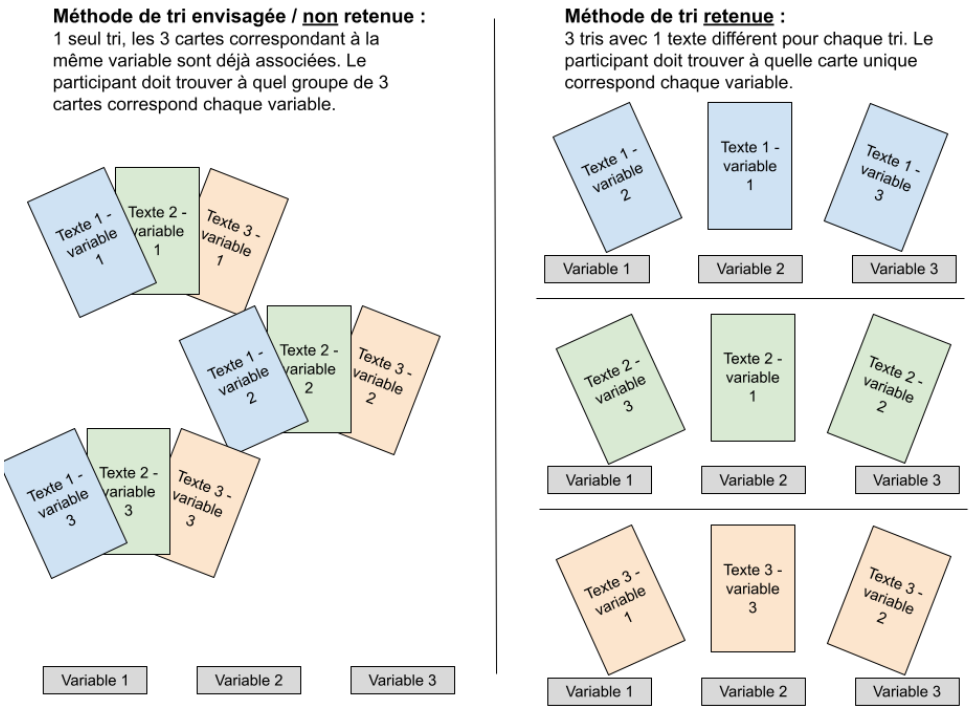

Figure 11 : Comparaison des deux options de tri de cartes de validation du matériel

Processus de tri :

Pour les 3 scénarios, les jeux de 12 cartes (facteurs retenus) étaient présentés dans un ordre différent à chaque participant.

Le participant devait associer chaque carte « message » au facteur et à la définition correspondante. Après chaque tri d'un jeu de 12 cartes, l'expérimentateur contrôlait les erreurs de classification (association d'une carte « message » avec une carte « définition » qui ne décrit pas le facteur qu'elle a été conçue pour représenter) et demandait au participant d'indiquer les cartes pour lesquelles il avait eu des doutes et pourquoi. En cas d'erreur, l'expérimentateur indiquait l'erreur au participant et cherchait à en comprendre l'origine avec lui.

Le but était d'évaluer si certaines cartes ou certains paragraphes représentaient moins bien que les autres le facteur associé.

À la fin de l'expérimentation, l'expérimentateur demandait au participant si les labels (facteurs) et leurs définitions lui paraissaient clairs.

3.4.3 Phase B1 : Lancement du tri de cartes en ligne

3.4.3.1 Recrutement

Le recrutement des participants pour le tri de cartes en ligne a été effectué auprès d'universitaires, d'étudiants, d'entreprises (THALES LAS France) et au sein du Ministère des Armées. L'objectif était d'avoir au moins 100 participants pour cette phase, afin d'assurer une diversité de profils et une robustesse statistique permettant d'identifier les facteurs les plus crédibles.

3.4.3.2 Outil de tri de cartes en ligne

Développement :

Figure 12 : Page « Tri de cartes » du site web créé pour l'expérimentation

Après avoir évalué les solutions commerciales disponibles pour mettre en œuvre notre tri de cartes, il s'est avéré qu'aucune d'entre elles ne permettait de « chaîner » trois tris de cartes consécutifs dans un ordre aléatoire. Nous avons donc développé un outil dédié, basé sur une solution open-source de GitHub développée par Stranovsky[5]. Cette solution ne permettant elle-même que de faire un seul tri de cartes, elle a dû être largement adaptée pour inclure : des pages d'introduction et questionnaires, trois tris de

[5] https://github.com/Luxato/Free-Cardsort

cartes consécutifs en ordre aléatoire, randomisation de l'ordre de présentation des cartes, limitation aux quatre catégories proposées, adaptation du design du site aux cartes choisies, traduction en français… Un aperçu du résultat final est donné en Figure 12.

Parcours du participant :

Le parcours du participant est illustré en Figure 13.

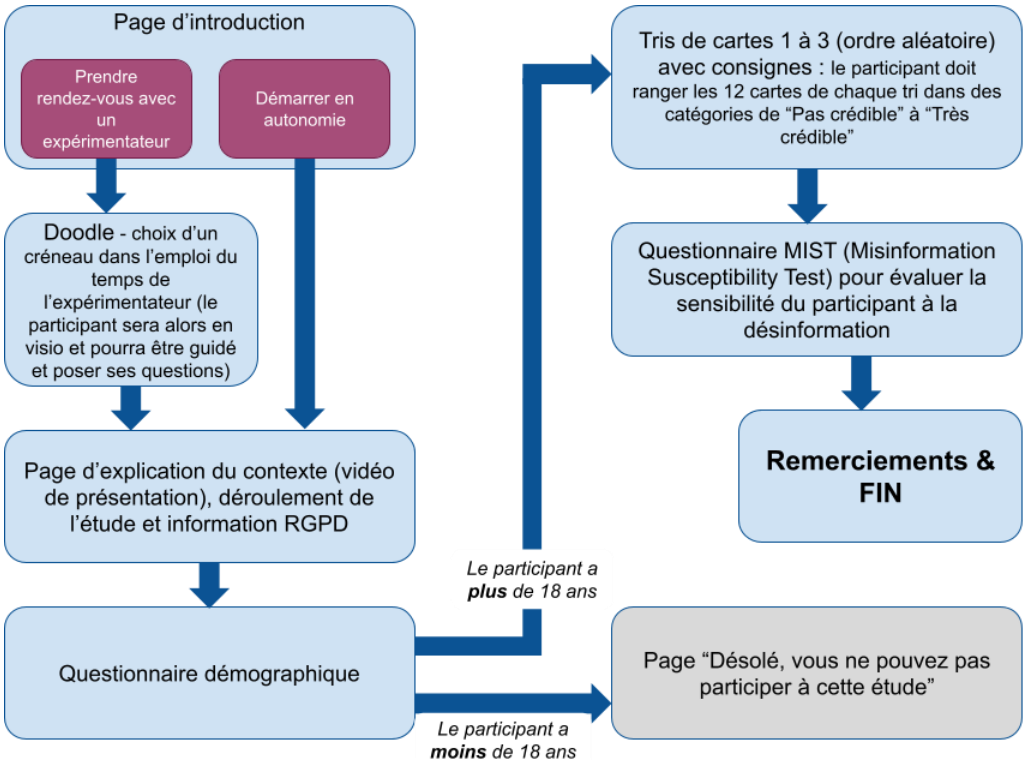

Figure 13 : Parcours du participant sur le site web de l'expérimentation

Après une mise en contexte vidéo et le questionnaire démographique, nous vérifions l'âge du participant afin d'exclure les mineurs pour des raisons de responsabilité. Le participant doit ensuite compléter trois tris de cartes successifs, chacun présentant les 12 variantes d'un des trois textes sélectionnés et validés lors des étapes précédentes. Pendant ces tris, le participant a pour consigne de classer chaque texte dans quatre catégories possibles : « Pas crédible », « Peu crédible », « Assez crédible » ou « Très crédible. Nous demandons également au participant de trier les items (ou textes) au sein de chaque boîte, du plus crédible au moins crédible. Au moment de l'enregistrement des réponses, un pop-up rappelle au participant de trier les cartes au sein de chaque catégorie avant de passer au tri suivant.

3.4.3.3 Envoi du tri de cartes aux participants

Tableau 11 : Répartition des modes d'administration du tri de cartes pour les participants de la phase B

Participants phase B	Avec un expérimentateur	En autonomie	Total
En ligne	8	267	275
En présentiel	2	0	2
Total	10	267	**277**

Le tri de cartes est destiné à des participants francophones et a été majoritairement réalisé en autonomie. Cependant, une dizaine de participants l'ont effectué en présence d'un expérimentateur, en présentiel ou en visioconférence (Tableau 11), ce qui a permis de recueillir des informations sur les difficultés rencontrées et les méthodes de tri utilisées.

3.4.4 Phase B2 : Tri de cartes en ligne

Le tri de cartes en ligne mis en œuvre pour les passations se déroule de la façon suivante :

1. Consignes et Contexte : présentation des consignes, du contexte de l'étude, et des informations RGPD (Règlement Général sur la Protection des Données).
2. Questionnaire Démographique
3. Tris : le participant effectue trois tris de cartes successifs, un pour chaque scénario. Lors de chaque tri, le participant est confronté aux 12 variations d'un seul scénario. L'ordre de présentation des scénarios n'est pas le même pour tous les participants, afin de compenser les effets d'entraînement et de fatigue cognitive.
4. Questionnaire MIST : évaluation de la sensibilité du participant aux fausses informations.

Le questionnaire MIST présenté est une version traduite en français du questionnaire proposé par l'Université de Cambridge (Maertens et al., 2023). Il est constitué de 20 intitulés d'articles divers, dont la moitié sont de vraies informations et l'autre moitié relèvent de la désinformation. Le participant doit catégoriser chaque titre d'article comme « Vrai » ou « Faux ». Ce questionnaire mesure la sensibilité aux fausses informations, les « faux positifs » indiquant une tendance à être crédule et les « faux négatifs » une méfiance excessive. Un score élevé indique une faible sensibilité à la désinformation.

L'ensemble de la passation dure environ 20 à 30 minutes.

3.5 Résultats de la phase A

Le tri de cartes au format papier, avec pour objectif d'associer chaque carte « message » avec la carte « définition » (soit le facteur de crédibilité) qu'elle représente, a été testé avec 21 participants, dont 9 femmes et 12 hommes. La majorité (12/21) possédait un diplôme de niveau Bac+5 et 15/21 avaient 30 ans ou moins. Les six ordres de passation possibles des trois tris consécutifs ont été appliqués de manière équilibrée, avec 3 à 5 participants par ordre.

3.5.1 Tri sur le jeu de cartes #1 « Pochoirs d'étoiles de David »

Pour ce jeu de cartes, 59 erreurs de classement (association d'une carte « message » avec une carte « définition » qui ne décrit pas le facteur qu'elle a été conçue pour représenter) ont été observées sur 252 classements, soit un taux d'erreur de 23,4%. La Figure 14 montre la répartition des erreurs. La carte « Détails » a été correctement classée seulement 6 fois sur 21, souvent confondue avec « Exemples ». De même, « Exemples » a été classée comme « Détails » aussi souvent que comme « Exemples » (10 fois chacun). « Répétitions » et « Pas clair » ont été bien classées respectivement 14 et 15 fois sur 21. Les autres cartes ont été correctement classées au moins 75% du temps, sans erreur pour « Nombres ».

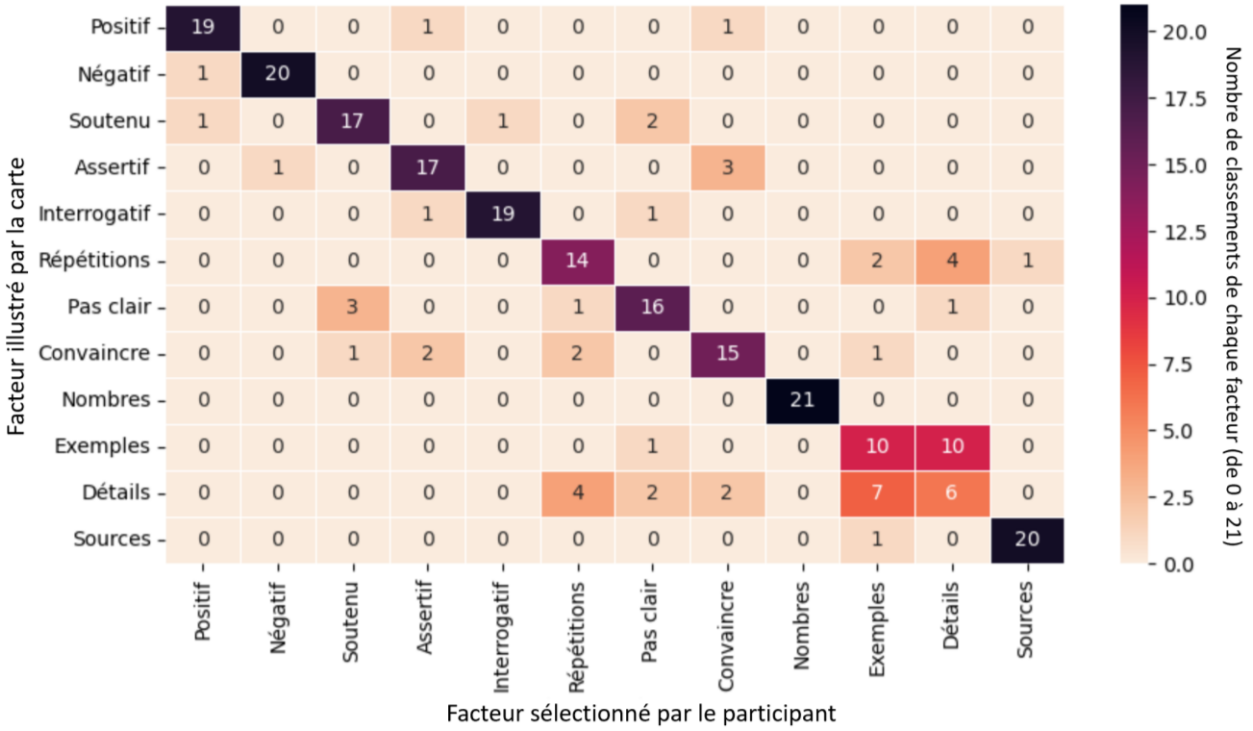

Figure 14 : Classement des cartes du scénario #1 « Pochoirs d'étoiles de David » par les participants pendant la phase A de l'expérimentation

3.5.2 Tri sur le jeu de cartes #2 « Algorithmes & élections »

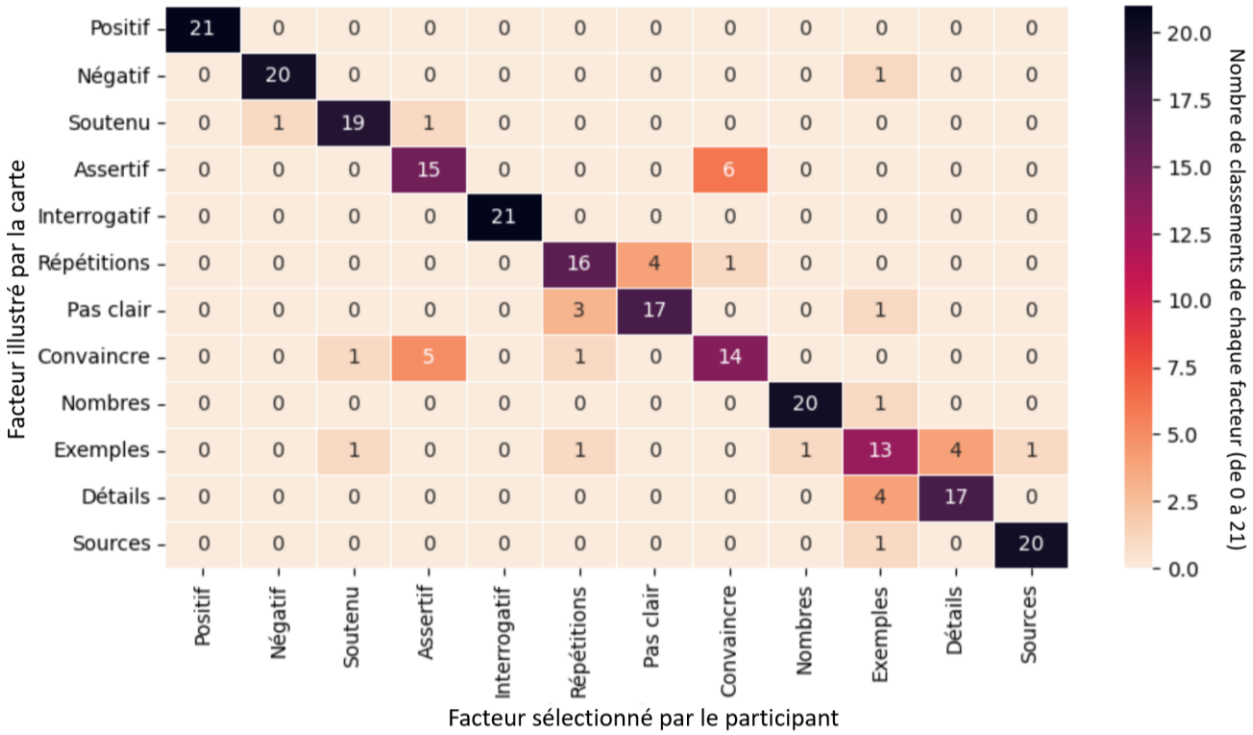

Figure 15 : Classement des cartes du scénario #2 « Algorithmes & élections » par les participants pendant la phase A de l'expérimentation

Pour ce jeu, nous avons enregistré 39 erreurs sur 252 classements, soit un taux d'erreur de 15,5%. La Figure 15 montre que la carte « Exemples » a été mal classée 8 fois, souvent confondue avec « Détails ».

Les cartes « Convaincre » et « Assertif » ont été fréquemment interverties. Une légère confusion a été observée entre « Répétitions » et « Pas clair ». Aucune erreur n'a été commise pour « Positif » et « Interrogatif ».

3.5.3 Tri sur le jeu #3 « Taylor Swift & Selena Gomez »

Avec 32 erreurs sur 252 classements (taux d'erreur de 12,7%), la Figure 16 montre que la carte « Convaincre » a été mal classée 8 fois, souvent comme « Négatif ». « Négatif » a également été classé comme « Convaincre » 3 fois. Les autres cartes ont été correctement classées au moins 80% du temps, avec des erreurs minimes pour « Exemples » et « Détails ». Aucune erreur n'a été observée pour « Interrogatif », « Nombres » et « Sources ».

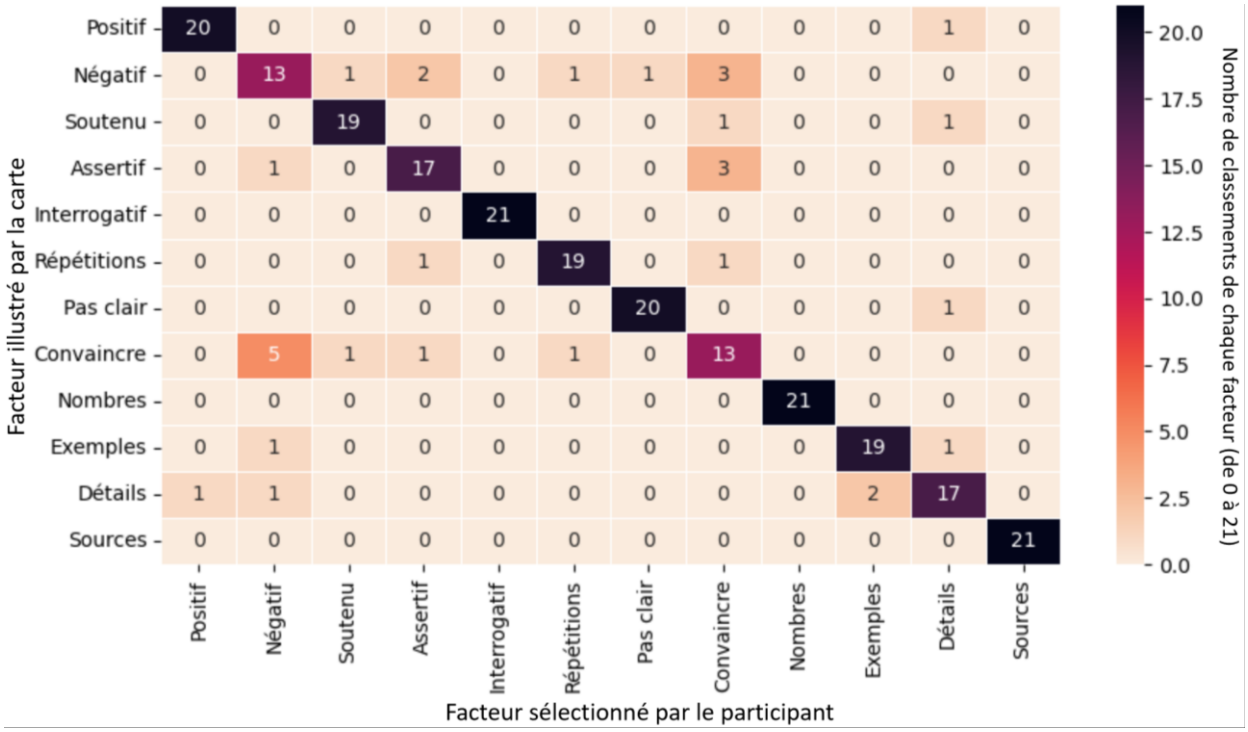

Figure 16 : Classement des cartes du scénario #3 « Taylor Swift & Selena Gomez » par les participants pendant la phase A de l'expérimentation

> ➢ **Bilan des résultats de la phase A**
>
> La plupart des cartes illustrent bien les facteurs associés.
>
> Cependant, les résultats révèlent une confusion élevée entre les cartes « Détails » et « Exemples » dans les deux premiers lots de cartes : ils pourront être rassemblés en un seul facteur dans la suite de la thèse (expérimentation 2). Le deuxième jeu présente également une confusion entre « Assertif » et « Convaincre ». Le troisième jeu introduit une confusion entre « Négatif » et « Convaincre ». Nous pouvons estimer que ces cartes ont été moins bien formulées pour illustrer les facteurs correspondants et doivent être traitées avec plus de prudence lors de la phase B.

L'analyse des verbatims a révélé les éléments suivants :

- La majorité des participants ont trouvé les différences entre les cartes claires après révélation des erreurs.
- Les définitions associées aux facteurs sont jugées claires par la majorité des participants.
- Un effet d'entraînement a été observé, les participants étant plus à l'aise dès le deuxième tri.
- Le lot de cartes « Pochoirs d'étoiles de David » a été perçu comme plus difficile, avec un taux d'erreurs plus élevé, notamment en raison de la confusion entre « Détails » et « Exemples ».
- Certains participants ont utilisé des indices visuels (ponctuation, chiffres) pour classer les cartes plus facilement sans lire le texte. Cela a pu rendre plus facile le tri de cartes telles que « Nombres », « Sources », « Interrogatif » et « Assertif ».

3.6 Résultats de la phase B

3.6.1 Caractéristiques des participants

L'étude de tri de cartes en ligne (phase B) a été diffusée auprès de divers groupes : étudiants de l'Université de Bordeaux, personnel et étudiants de l'École Nationale Supérieure de Cognitique (ENSC, Bordeaux INP), collaborateurs THALES, et personnels du Ministère des Armées. Il a également été diffusé via des canaux tels que LinkedIn afin de récolter des réponses plus diversifiées. Les participants de la phase A n'ont pas été inclus dans la phase B pour éviter qu'ils reconnaissent les facteurs liés aux cartes.

Au total, 695 personnes ont répondu, mais seulement 277 ont complété l'étude, produisant des résultats exploitables. Parmi elles, 128 hommes, 146 femmes, et 3 personnes dont le genre n'est pas identifié. Plus de la moitié des participants sont des étudiants de l'Université de Bordeaux (voir Tableau 12), avec d'autres populations significatives provenant du Ministère des Armées, de l'ENSC, et de THALES. Tous les participants sauf 7 sont de nationalité française. 74% des participants ont entre 18 et 30 ans et seulement 2% ont 65 ans ou plus.

Tableau 12 : Répartition des participants par entité de rattachement

Entité de rattachement	Nombre
Université de Bordeaux (étudiants)	149
Ministère des Armées	33
ENSC (étudiants)	26
Autres écoles, universités & centres de recherche	19
THALES	19
Aucune	16
Autres entreprises	11
Autres institutions publiques	4

Les niveaux d'études des participants sont distribués de Bac à Bac+8 (Figure 17), avec un pic au niveau Bac et un autre à Bac+5 qui peuvent s'expliquer par la diffusion de l'étude aux étudiants de l'Université

de Bordeaux (dont de nombreux étudiants entre L1 et L3 qui n'ont donc pas encore obtenu un diplôme autre que le baccalauréat) et à des professionnels (Ministère des Armées, THALES).

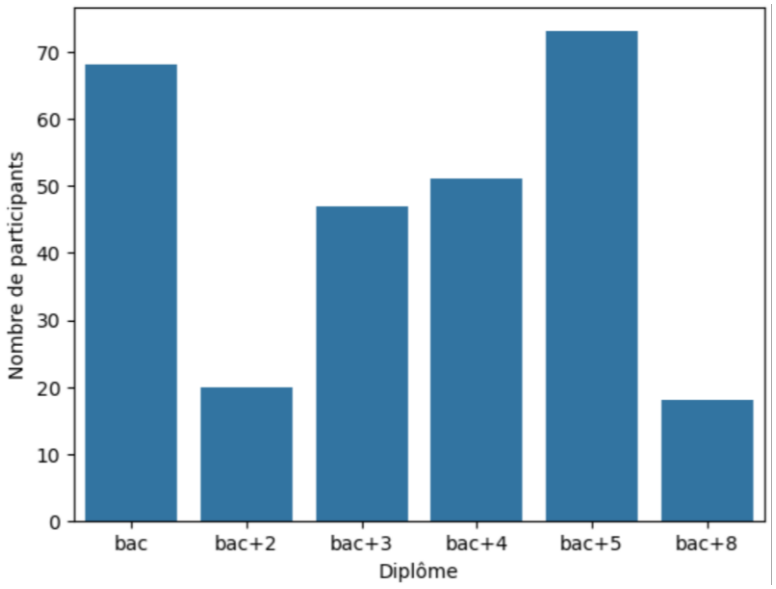

Figure 17 : Distribution des participants par niveau d'études

3.6.1.1 Durée de passation de la phase B

La durée moyenne de passation de l'expérimentation est de 21 minutes, avec des durées allant de 3 minutes à 1h06 (Figure 18). Les durées supérieures à 1h20 ont été exclues pour éviter les biais liés aux pauses. Les durées les plus courtes peuvent soulever des questions quant au sérieux des participants concernés. Cependant, nous pouvons argumenter que ces participants consomment l'information de manière rapide, en diagonale, et ce sont justement leurs conditions de consommation de l'information habituelles que nous cherchons à évaluer dans cette étude.

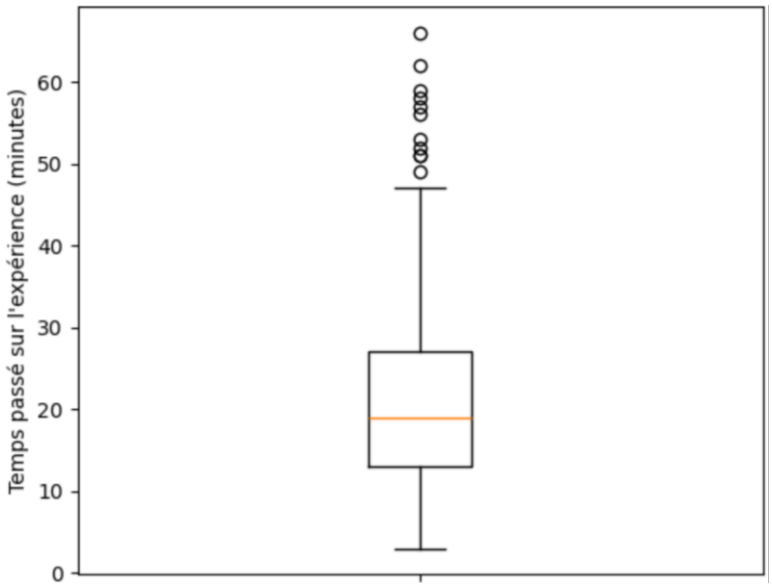

Figure 18 : Distribution du temps de passation de l'expérimentation (phase B)

3.6.1.2 Analyse factorielle de données mixtes des caractéristiques des participants

Une Analyse Factorielle des Données Mixtes (AFDM) (Pagès, 2014) a été réalisée sur les caractéristiques des participants : genre, âge, diplôme, entité de rattachement, score MIST et scores de faux positifs/négatifs. Le diplôme a été transformé en variable numérique (nombre d'années d'études après le baccalauréat).

Hypothèse testée : Existe-t-il une relation entre certaines caractéristiques des participants ?

Les résultats montrent une forte corrélation entre le score MIST et les scores de faux positifs/négatifs, conduisant à ne conserver que le score MIST dans le modèle final. Les deux premières dimensions de l'AFDM obtenue (Figure 19) capturent seulement 29,37% de la variance. Elle semble cependant indiquer une relation entre l'entité de rattachement et le groupe d'âge, qui pourrait s'expliquer par la différenciation entre les entités composées d'étudiants (Université de Bordeaux et École Nationale Supérieure de Cognitique) et les entités de rattachement de divers professionnels (voir Tableau 12).

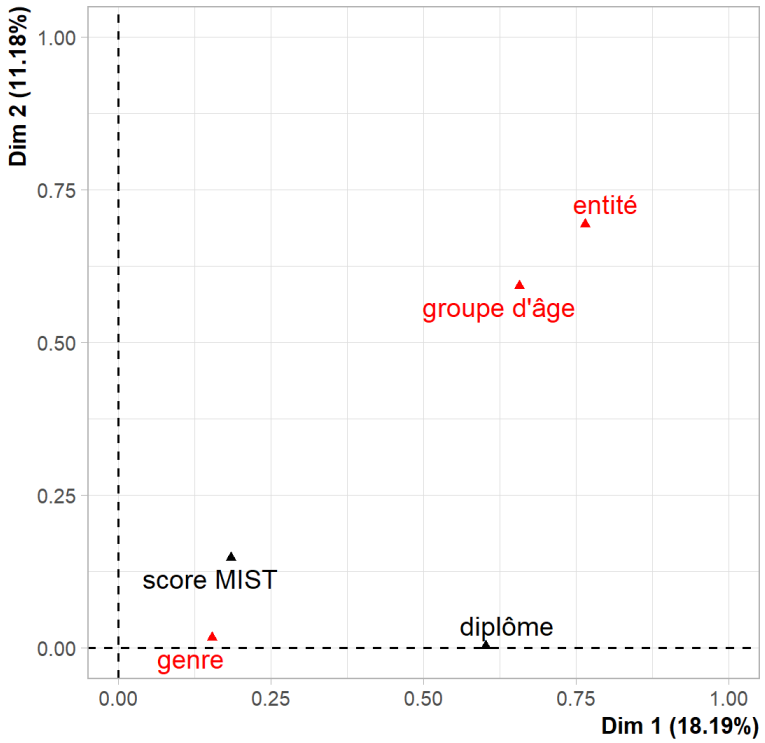

Figure 19 : Graphe des caractéristiques des participants suivant les dimensions 1 et 2 de l'AFDM

Dans les dimensions observées, le score MIST semble également avoir une corrélation avec le genre. La répartition des notes par genre (Figure 20) montre en effet que les femmes ont un score MIST moyen légèrement plus bas que celui des hommes dans l'échantillon de l'étude (14,6/20 vs 15,4/20). Ce phénomène pourrait s'expliquer par la présence de plus d'hommes parmi les participants interrogés ayant les plus hauts niveaux d'études (bac+5 et bac+8).

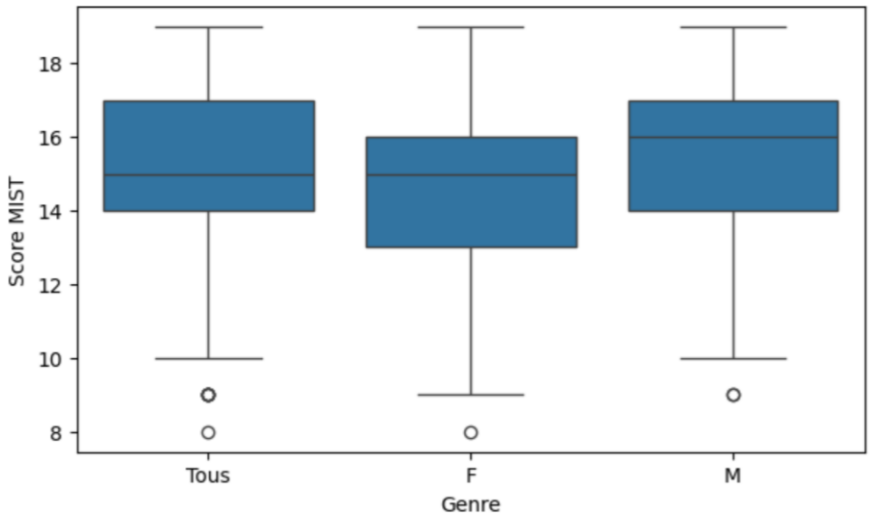

Figure 20 : Répartition du score MIST en fonction du genre

3.6.1.3 Niveau d'études et score MIST

Le score MIST moyen est de 14,97, le score médian est de 15, et seulement 2% des participants ont eu moins de 10/20. D'après les scores MIST obtenus, parmi les 277 participants, 66% ont tendance à être méfiants envers l'information (plus de faux négatifs que de faux positifs), 17% ont tendance à être crédules (plus de faux positifs que de faux négatifs) et 17% ont des résultats équilibrés (autant de faux positifs que de faux négatifs). Le nombre de faux positifs médian est de 1 (moyenne 1,60) et le nombre de faux négatifs médian est de 3 (moyenne 3,08).

Tableau 13 : Comparaison des scores MIST, faux positifs et faux négatifs entre notre étude et celle de Cambridge à l'origine du score MIST (Maertens et al., 2023)

	Notre étude	Univ. Cambridge étude 1	Univ. Cambridge étude 2
Population	277 français	409 états-uniens	3479 états-uniens
Score MIST	Médiane : 15 Moyenne : 14,97 Écart-type : 2,34	Moyenne : 15,71 Écart-type : 3,35	Médiane : 14
Nb faux positifs (« naïf »)	Médiane : 1 Moyenne : 1,60 Écart-type : 1,66	Moyenne : 1,91 Écart-type : 2,10	Médiane : 2
Nb faux négatifs (« sceptique »)	Médiane : 3 Moyenne : 3,08 Écart-type : 1,86	Moyenne : 2,38 Écart-type : 2,43	Médiane : 3

Le Tableau 13 propose une comparaison du score MIST obtenu avec les études menées par l'Université de Cambridge (Maertens et al., 2023) : s'il existe une différence entre les résultats de notre expérimentation et ceux des études de développement (étude 1) et de validation (étude 2) de l'Université de Cambridge, le nombre de faux négatifs reste toujours plus élevé que le nombre de faux positifs.

Hypothèse testée : Existe-t-il une relation entre le niveau d'études et le score MIST des participants ?

Un test de Pearson (Freedman et al., 2007) révèle une corrélation légère (22,2%) entre le niveau d'étude et le score MIST (Figure 21). La p-value inférieure à 0,01 confirme que les variables ne sont pas

indépendantes. Nous en déduisons que parmi les populations évaluées (baccalauréat au minium), les personnes ayant un niveau d'études plus élevé ont tendance à mieux distinguer les informations vraies des informations fausses.

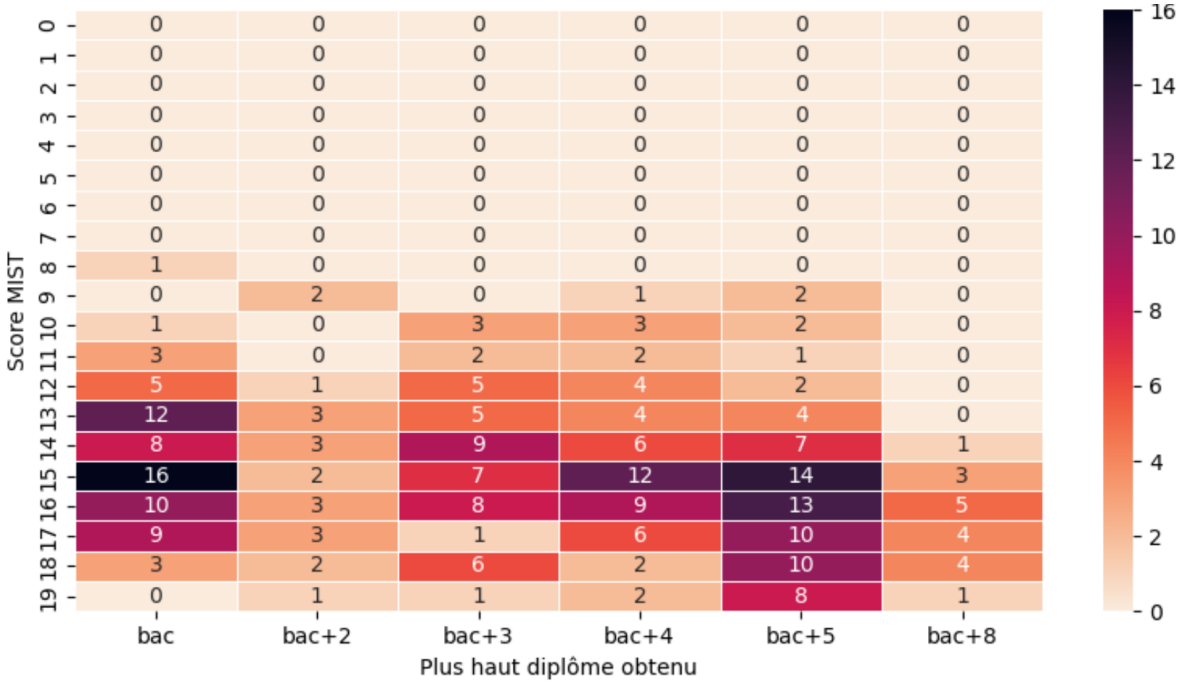

Figure 21 : Score MIST obtenu en fonction du diplôme (nombre de participants)

3.6.2 Crédibilité des facteurs

Dans ce chapitre, nous considérons que le classement des facteurs de crédibilité réalisé par les participants dans des « boîtes » de niveaux de crédibilité s'apparente à une échelle de type Likert. Nous faisons ainsi l'hypothèse d'une distance équivalente entre chaque modalité de réponse, ce qui permet de convertir les jugements qualitatifs en valeurs numériques : « pas crédible » = 1, « assez crédible » = 2, « peu crédible » = 3 et « très crédible » = 4.

Par ailleurs, les classements effectués pour chaque scénario provenant des mêmes participants, les échantillons sont considérés comme dépendants. Pour vérifier la normalité des échantillons, nous avons utilisé le test de Shapiro-Wilk. Les résultats indiquent que les échantillons ne suivent pas une distribution normale (p-value < 0,01) (Shapiro & Wilk, 1965).

Hypothèses testées :

1) Existe-t-il une différence significative entre les facteurs classés comme les plus crédibles et ceux classés comme les moins crédibles ?
2) Entre quels facteurs spécifiques observe-t-on une différence significative ?
3) Y a-t-il une différence entre les scénarios ?

Pour tester ces hypothèses, nous avons choisi d'utiliser le test non paramétrique de Friedman (1937), qui est adapté à des données ordinales et non normales et permet de comparer plus de deux groupes. Ce test évalue l'hypothèse nulle (H0) selon laquelle il n'existe pas de différence significative entre les groupes. Si cette hypothèse est rejetée, nous appliquons un test post-hoc de Nemenyi (1963) pour identifier précisément les facteurs entre lesquels les différences sont significatives.

a) Scénario 1 :

Le boxplot de la distribution des valeurs de crédibilité (Figure 22) révèle que dans le premier scénario (les pochoirs d'étoiles de David), certains facteurs sont plus fréquemment classés comme peu ou pas crédibles (comme *positif* et *convaincre*), alors que d'autres sont plus souvent classés comme crédibles voire très crédibles (comme la présence d'*exemples*, de *détails* ou de *sources*).

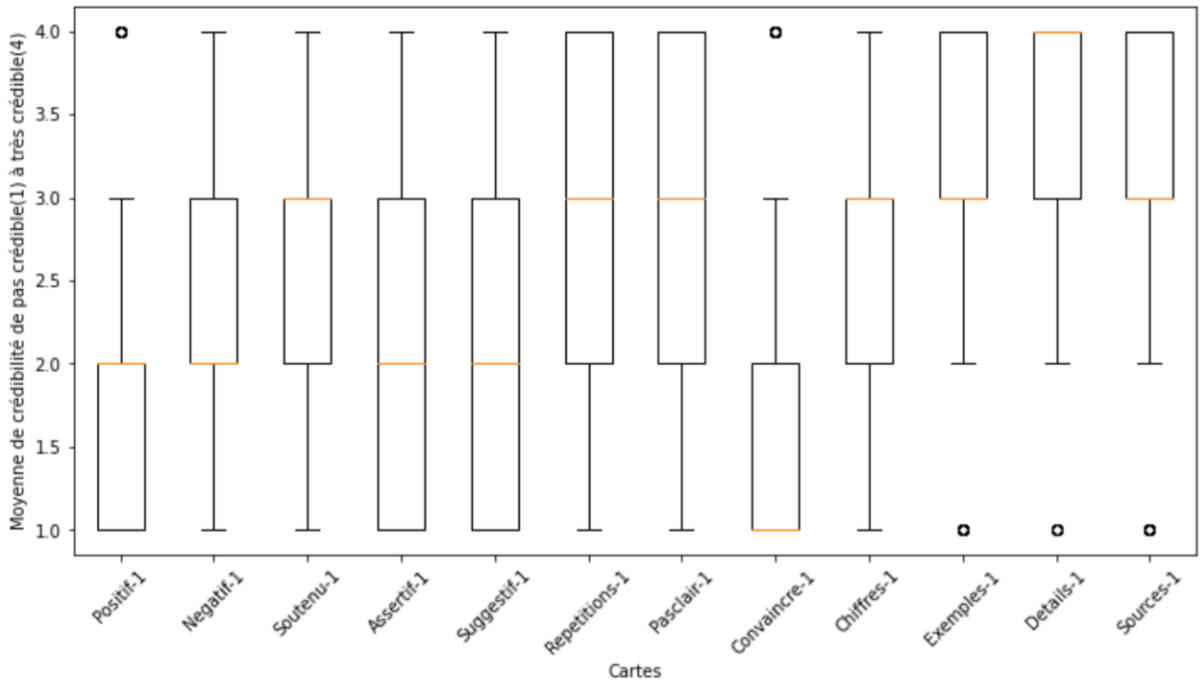

Figure 22 : Valeurs de crédibilité pour chaque facteur pour le scénario 1 (pochoirs)

Le test non paramétrique de Friedman confirme qu'il existe des différences significatives entre les 12 facteurs, avec une p-value inférieure au seuil de 0,01 qui nous permet de rejeter l'hypothèse nulle. Nous appliquons alors un test post-hoc de Nemenyi (1963) en ordonnant les facteurs par crédibilité moyenne.

Le test post-hoc de Nemenyi (Figure 23) montre une différence significative de moyenne de crédibilité entre les facteurs les moins crédibles et les plus crédibles. Cependant, aucune différence significative n'est observée entre les facteurs les moins crédibles (à savoir *convaincre*, *positif* et *assertif* pour le scénario 1). De même, parmi les facteurs les plus crédibles (*pas clair*, *répétitions*, *sources*, *exemples*), seul le facteur *détails* se distingue en étant significativement plus crédible que tous les autres.

	Convaincre	Positif	Assertif	Suggestif	Négatif	Chiffres	Soutenu	Pas clair	Répétitions	Sources	Exemples	Détails
Convaincre	1.000											
Positif	1.000	1.000										
Assertif	0.859	0.996	1.000									
Suggestif	0.002**	0.021*	0.364	1.000								
Négatif	0.001**	0.018*	0.337	1.000	1.000							
Chiffres	0.000**	0.000**	0.000**	0.009**	0.011*	1.000						
Soutenu	0.000**	0.000**	0.000**	0.000**	0.000**	0.971	1.000					
Pas clair	0.000**	0.000**	0.000**	0.000**	0.000**	0.004**	0.257	1.000				
Répétitions	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.002**	0.951	1.000			
Sources	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.735	1.000	1.000		
Exemples	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.337	0.997	1.000	1.000	
Détails	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.002**	0.018*	1.000

Figure 23 : Test post-hoc de Nemenyi pour le scénario 1 (pochoirs), avec des facteurs ordonnés par crédibilité moyenne (en rouge les facteurs les moins crédibles, en jaune l'ensemble des facteurs plutôt crédibles, en vert le facteur le plus crédible)

b) Scénario 2 :

Le boxplot de la distribution des valeurs de crédibilité (Figure 24) révèle que dans, le deuxième scénario (les algorithmes), certains facteurs sont plus fréquemment classés comme peu ou pas crédibles (comme les facteurs émotionnels, l'assertivité et la volonté de convaincre), alors que d'autres sont plus souvent classés comme crédibles voire très crédibles (comme la présence de détails ou de sources ou un message pas clair).

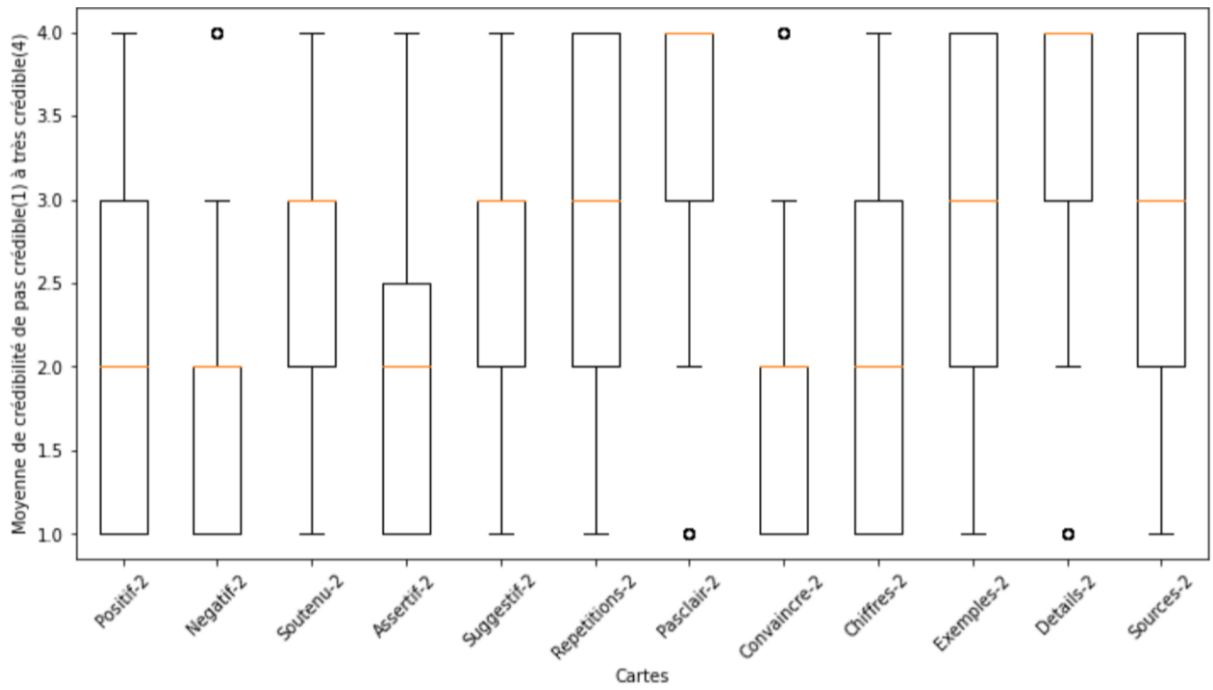

Figure 24 : Valeurs de crédibilité pour chaque facteur pour le scénario 2 (algorithmes)

Comme pour le scénario précédent, le test non paramétrique de Friedman confirme qu'il existe des différences significatives entre les 12 facteurs. Nous appliquons donc un test post-hoc de Nemenyi (1963).

Celui-ci (Figure 25) montre une différence significative de moyenne de crédibilité entre les facteurs les moins crédibles et les plus crédibles. Cependant, il n'y a pas de différence significative entre les facteurs les moins crédibles (qui sont *négatif*, *assertif*, *convaincre*, *positif* et *chiffres* pour le scénario 2), ni entre les facteurs les plus crédibles (*pas clair* et *détails*).

	Négatif	Assertif	Convaincre	Positif	Chiffres	Suggestif	Soutenu	Exemples	Sources	Répétitions	Pas clair	Détails
Négatif	1.000											
Assertif	1.000	1.000										
Convaincre	0.991	1.000	1.000									
Positif	0.254	0.633	0.949	1.000								
Chiffres	0.254	0.633	0.949	1.000	1.000							
Suggestif	0.000**	0.000**	0.000**	0.000**	0.000**	1.000						
Soutenu	0.000**	0.000**	0.000**	0.000**	0.000**	0.787	1.000					
Exemples	0.000**	0.000**	0.000**	0.000**	0.000**	0.056	0.969	1.000				
Sources	0.000**	0.000**	0.000**	0.000**	0.000**	0.032*	0.927	1.000	1.000			
Répétitions	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.254	0.983	0.995	1.000		
Pas clair	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.028*	0.050*	0.569	1.000	
Détails	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.001**	0.038*	0.990	1.000

Figure 25 : Test post-hoc de Nemenyi pour le scénario 2 (algorithmes), avec des facteurs ordonnés par crédibilité moyenne (en rouge les facteurs les moins crédibles, en vert les facteurs les plus crédibles)

c) Scénario 3 :

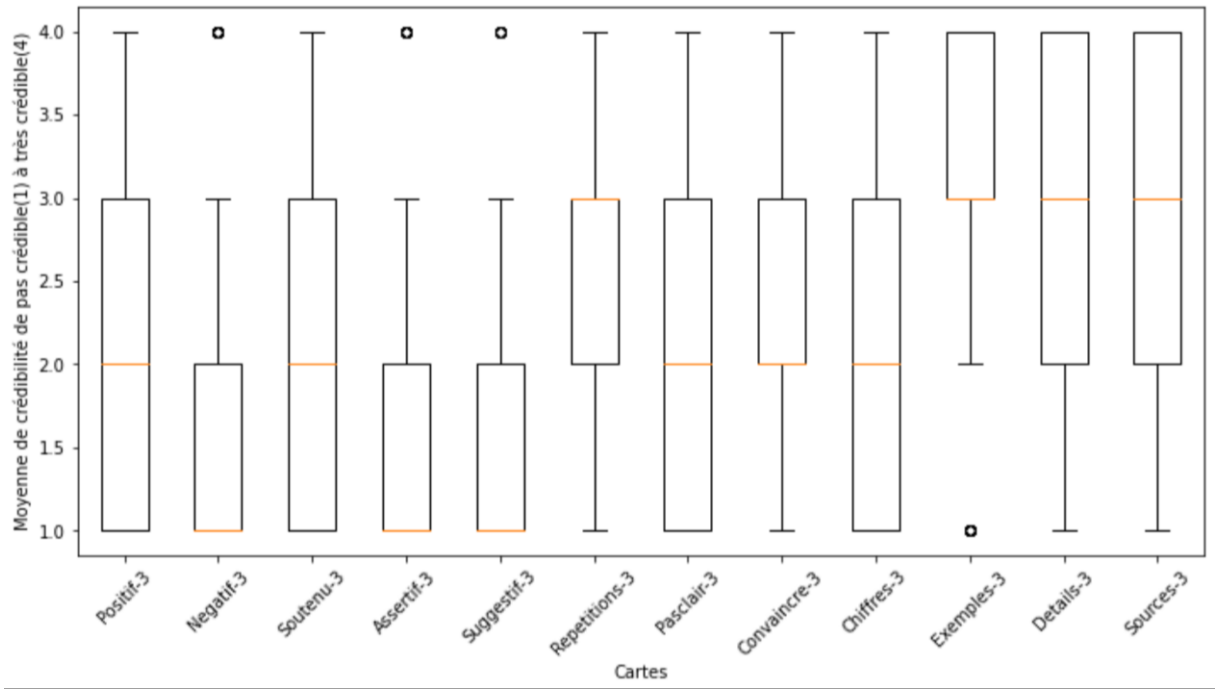

Figure 26 : Valeurs de crédibilité pour chaque facteur pour le scénario 3 (Taylor Swift & Selena Gomez)

Le boxplot de la distribution des valeurs de crédibilité (Figure 26) révèle que dans, le troisième scénario (Taylor Swift & Selena Gomez), certains facteurs sont plus fréquemment classés comme peu ou pas crédibles (comme les facteurs émotionnels, l'assertivité ou la suggestion), alors que d'autres sont plus souvent classés comme crédibles voire très crédibles (comme la présence d'*exemples*, de *détails* ou de *sources*).

Comme pour les scénarios précédents, le test non paramétrique de Friedman confirme qu'il existe des différences significatives entre les 12 facteurs. Nous appliquons donc un test post-hoc de Nemenyi (1963). Celui-ci (Figure 27) montre une différence significative de moyenne de crédibilité entre les facteurs les moins crédibles et les plus crédibles. Cependant, il n'y a pas de différence significative entre les facteurs les moins crédibles (qui sont *suggestif* et *assertif* pour le scénario 3), ni entre les facteurs les plus crédibles (*sources* et *exemples*).

	Suggestif	Assertif	Négatif	Chiffres	Positif	Pas clair	Soutenu	Convaincre	Répétitions	Détails	Sources	Exemples
Suggestif	1.000											
Assertif	0.995	1.000										
Négatif	0.004**	0.143	1.000									
Chiffres	0.000**	0.000**	0.403	1.000								
Positif	0.000**	0.000**	0.330	1.000	1.000							
Pas clair	0.000**	0.000**	0.221	1.000	1.000	1.000						
Soutenu	0.000**	0.000**	0.000**	0.590	0.671	0.797	1.000					
Convaincre	0.000**	0.000**	0.000**	0.037*	0.052	0.093	0.988	1.000				
Répétitions	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.002**	0.133	1.000			
Détails	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.002**	0.983	1.000		
Sources	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.053	1.000	
Exemples	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.495	1.000

Figure 27 : Test post-hoc de Nemenyi pour le scénario 3 (Taylor Swift & Selena Gomez), avec des facteurs ordonnés par crédibilité moyenne (en rouge les facteurs les moins crédibles, en jaune les facteurs plutôt crédibles, en vert les facteurs les plus crédibles)

d) Moyennes de crédibilité pour les 3 scénarios confondus :

Nous avons précédemment constaté que les trois scénarios ne présentent pas toujours les mêmes facteurs comme étant les plus ou les moins crédibles. La Figure 28 montre également que le troisième scénario (Swift & Gomez) est globalement perçu comme moins crédible que les autres. La Figure 30 montre les facteurs ordonnés du moins au plus crédible, pour les 3 scénarios confondus.

Moyenne de crédibilité sur les scénarios :

Les moyennes de crédibilité pour les scénarios 1 et 2 sont respectivement de 2,52 et 2,51, toutes deux proches de la moyenne entre « peu crédible » et « assez crédible ». En revanche, le scénario 3 affiche une moyenne de 2,21, plus proche de « peu crédible ». Un test statistique de Friedman confirme une différence significative entre les moyennes de crédibilité des scénarios (p-value < 0,01).

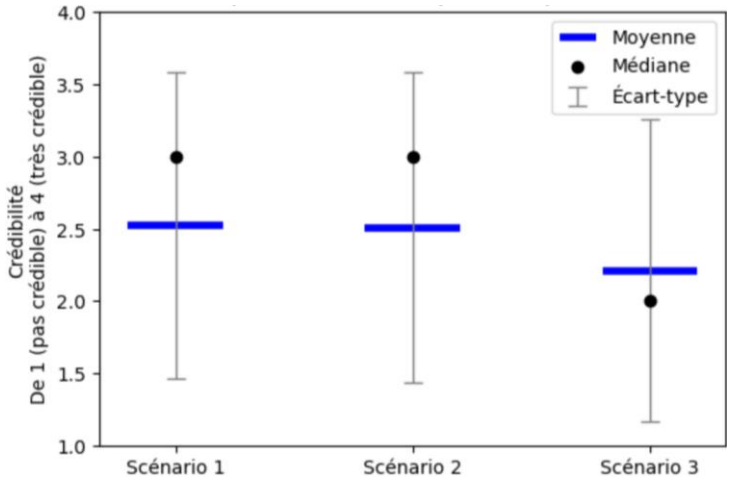

Figure 28 : Moyenne de crédibilité obtenue pour chaque scénario

Pour identifier précisément les différences, nous effectuons un test post-hoc de Nemenyi (Figure 29). La p-value est inférieure au seuil de significativité de 0,01 uniquement entre le scénario 3 et les deux autres. Ainsi, nous concluons que les scénarios 1 et 2 ne diffèrent pas significativement en termes de crédibilité, tandis que le scénario 3 (qui relève de la désinformation pure) est considéré comme significativement moins crédible que les deux autres par les participants.

	Scenario 1	**Scenario 2**	**Scenario 3**
Scenario 1	1.000		
Scenario 2	0.984	1.000	
Scenario 3	0.000**	0.000**	1.000

Figure 29 : Test post-hoc de Nemenyi pour les moyennes globales des 3 scénarios (en rouge le scénario le moins crédible)

Moyenne de crédibilité sur les facteurs :

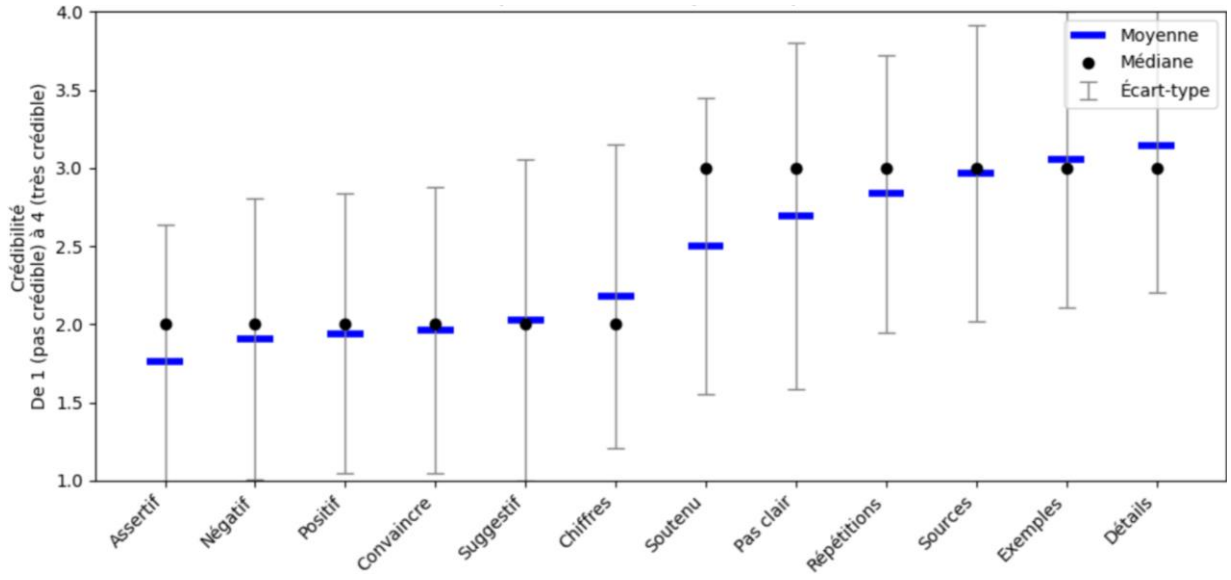

Figure 30 : Moyenne de crédibilité obtenue pour chaque facteur, tous scénarios confondus

Pour les 3 scénarios combinés, le test non paramétrique de Friedman donne à nouveau une p-value inférieure au seuil de 0,01. Cela confirme qu'il existe une différence significative entre les 12 facteurs. Pour identifier précisément quels facteurs diffèrent, nous avons effectué un test post-hoc de Nemenyi. Celui-ci (Figure 31) montre une différence significative entre la moyenne de crédibilité associée aux facteurs les moins crédibles et les plus crédibles. Cependant, parmi les facteurs les moins crédibles (*assertif*, *négatif*, *positif* et *convaincre* pour les 3 scénarios combinés), seul le facteur *assertif* est perçu comme significativement moins crédible que les autres. Il n'y a pas non plus de différence significative entre les deux facteurs les plus crédibles (*exemples* et *détails*), ni entre *exemples* et *sources*, qui font également partie des facteurs les plus crédibles.

	Assertif	Négatif	Positif	Convaincre	Suggestif	Chiffres	Soutenu	Pas clair	Répétitions	Sources	Exemples	Détails
Assertif	1.000											
Négatif	0.183	1.000										
Positif	0.011*	0.999	1.000									
Convaincre	0.001**	0.919	1.000	1.000								
Suggestif	0.000**	0.492	0.971	1.000	1.000							
Chiffres	0.000**	0.000**	0.002**	0.024*	0.186	1.000						
Soutenu	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	1.000					
Pas clair	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.054	1.000				
Répétitions	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.097	1.000			
Sources	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.514	1.000		
Exemples	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.008**	0.930	1.000	
Détails	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.086	0.927	1.000

Figure 31 : Test post-hoc de Nemenyi pour les 3 scénarios confondus, avec des facteurs ordonnés par crédibilité moyenne (en rouge le facteur le moins crédible, en orange les facteurs peu crédibles, en jaune les facteurs plutôt crédibles, en vert les facteurs les plus crédibles)

Classification des facteurs en fonction de leur niveau de crédibilité, pour les 3 scénarios confondus :

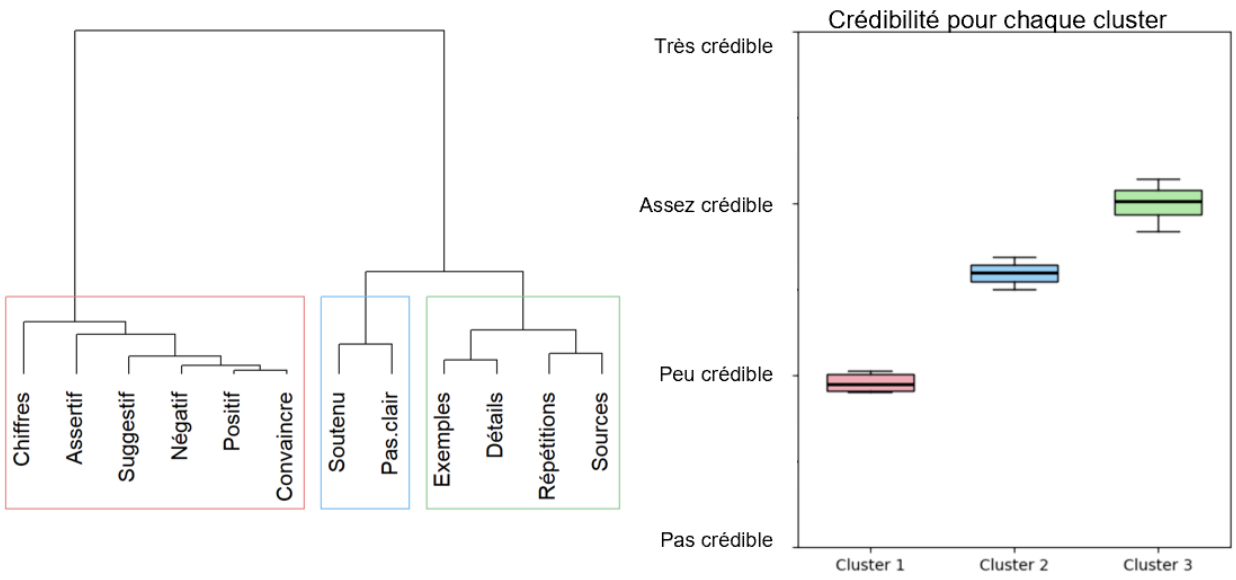

Figure 32 : Clusters en fonction des moyennes de crédibilité, par classification hiérarchique ascendante avec critère de Ward, appliquée aux distances entre moyennes de crédibilité

Le package *hclust* dans R nous permet de générer un dendrogramme des clusters des facteurs en fonction de leurs moyennes de crédibilité, par classification hiérarchique ascendante avec critère de Ward, appliquée aux distances entre moyennes de crédibilité (Figure 32). Le cluster 1 (*chiffres, assertif, suggestif, négatif, positif, convaincre*) regroupe les facteurs jugés en moyenne « peu crédibles ». Le cluster 2 (*soutenu, pas clair*) comprend les facteurs à la crédibilité neutre en moyenne. Enfin, le cluster 1 (*exemples, détails, répétitions, sources*) contient les facteurs en moyenne « assez crédibles ».

3.6.3 Rapport entre la crédibilité évaluée et les caractéristiques des participants

Dans ce paragraphe, nous comparons les caractéristiques (âge, niveau d'étude, genre, score MIST, etc.) des participants. Pour faciliter l'analyse, nous les avons divisés en deux groupes pour chaque caractéristique étudiée (par exemple, un groupe avec les scores MIST en dessous de la moyenne et un groupe au-dessus).

Hypothèses testées :

1 La crédibilité moyenne accordée par les participants varie-t-elle de manière significative en fonction de leurs caractéristiques individuelles ?
2 Les classements effectués par les participants sur les facteurs varient-ils de manière significative en fonction de leurs caractéristiques ?

La première hypothèse est évaluée à l'aide du test t de Student, qui est approprié pour comparer les moyennes de deux groupes indépendants. Ce test est robuste aux écarts par rapport à la normalité, surtout lorsque les tailles d'échantillon sont suffisamment grandes (généralement supérieures à 30) (Box, 1953). Dans notre étude, les groupes sont indépendants, car les valeurs de crédibilité sont attribuées par des participants différents. Le test évalue l'hypothèse nulle (H0) selon laquelle il n'existe pas de différence entre les moyennes des deux groupes.

Étant donné que les données ne suivent pas une distribution normale, nous avons utilisé le test non paramétrique de Mann-Whitney pour comparer les résultats des deux groupes sur chaque facteur. Ce test a été choisi car il ne repose pas sur l'hypothèse de normalité et permet d'évaluer l'hypothèse nulle (H0) selon laquelle la distribution des 2 populations est statistiquement similaire.

3.6.3.1 Rapport entre crédibilité des facteurs et score MIST du participant

Pour cette étude, nous avons divisé la population en deux catégories : score MIST faible et score MIST élevé. La moyenne et la médiane du score MIST dans cette étude sont de 15/20, ce sera donc le seuil en-deçà duquel le score sera considéré comme faible et au-delà fort.

Crédibilité moyenne :

Les participants ayant un score MIST plus faible donnent aux textes une crédibilité moyenne légèrement plus élevée (2,43 vs 2,39). Cependant, un test t de Student montre que ces différences ne sont pas statistiquement significatives (p-values proches mais supérieures au seuil de 0,05) : en conséquence, il faudrait étudier une population plus nombreuse pour confirmer cette observation.

Crédibilité selon les facteurs :

La Figure 33 montre que certains facteurs semblent être classés différemment selon le score MIST des participants.

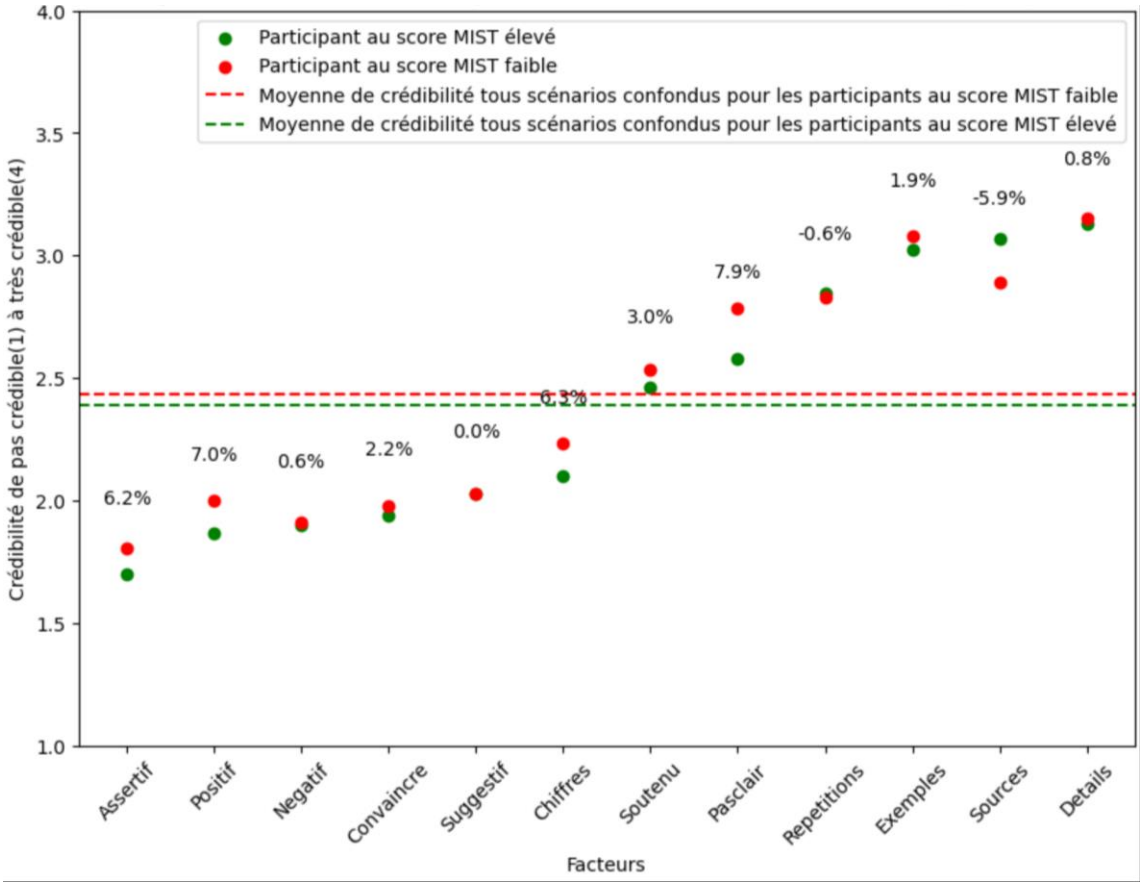

Figure 33 : Comparaison des moyennes de crédibilité pour chaque facteur pour les participants ayant un score MIST faible et élevé, avec un seuil à 15/20

Nous avons réalisé un test non paramétrique de Mann-Whitney pour comparer les résultats des deux groupes. La Figure 34 montre que l'hypothèse nulle est rejetée pour les facteurs *positif*, *pas clair* et *sources*. Les facteurs en question sont donc classés de façon significativement différente en fonction du score MIST.

	Positif	Négatif	Soutenu	Assertif	Suggestif	Répétitions	Pas clair	Convaincre	Chiffres	Exemples	Détails	Sources
U	91390.000	85622.000	88623.500	90421.500	84943.500	84039.000	93487.500	86533.500	90965.500	87484.500	86672.500	75826.500
p	0.046*	0.833	0.262	0.083	1.000	0.781	0.010**	0.624	0.067	0.433	0.589	0.005**

Figure 34 : Résultats du test de Mann-Whitney comparant les scores de crédibilité donnés par les participants ayant un score MIST faible et élevé, avec un seuil à 15/20

Ainsi, pour les 3 scénarios combinés, les participants ayant un score MIST élevé ont davantage valorisé la présence de *sources*, qu'ils jugent plus crédibles, et accordent moins de confiance aux *émotions positives* et à un langage *pas clair*.

3.6.3.2 Rapport entre crédibilité des facteurs et niveau d'études du participant

Pour comparer deux groupes équilibrés, nous avons fixé la limite entre un niveau d'études faible et élevé à Bac+3, incluant ce niveau dans les niveaux d'études faibles. Ainsi, 142 participants ont un niveau d'études élevé et 133 participants ont un niveau d'études plus faible. À nouveau, un test t de Student nous permet de conclure qu'il n'y a pas de différence significative entre les moyennes de crédibilité accordées quel que soit le groupe.

	Positif	Négatif	Soutenu	Assertif	Suggestif	Répétitions	Pas clair	Convaincre	Chiffres	Exemples	Détails	Sources
U	89401.500	87483.000	90390.500	83595.000	81361.500	83413.000	94830.500	82812.000	81369.500	92546.000	89925.000	79081.000
p	0.336	0.708	0.212	0.402	0.135	0.384	0.010*	0.291	0.139	0.054	0.257	0.029*

Figure 35 : Résultats du test de Mann-Whitney comparant les scores de crédibilité donnés par les participants ayant un niveau d'études faible (<=Bac+3) vs élevé

Le test de Mann-Whitney (Figure 35) nous permet alors de rejeter l'hypothèse nulle suivant laquelle les moyennes des 2 populations sont statistiquement similaires pour les facteurs *pas clair* et *sources* (p-values < 0,05). En effet, les messages *pas clairs* sont classés comme plus crédibles par les participants ayant un niveau d'études plus faible, et les messages avec présence de *sources* sont classés comme plus crédibles par les participants ayant un niveau d'études plus élevé.

3.6.3.3 Rapport entre crédibilité évaluée et persévérance du participant

Dans ce paragraphe, nous prenons en compte des participants qui ont été écartés jusqu'ici : ceux qui n'ont pas terminé l'expérimentation. Nous considérons ceux qui ont complété au moins un tri.

La Figure 36 montre que les participants qui ont abandonné le test ont globalement des niveaux d'études plus faibles que ceux qui ont persévéré et terminé l'expérimentation.

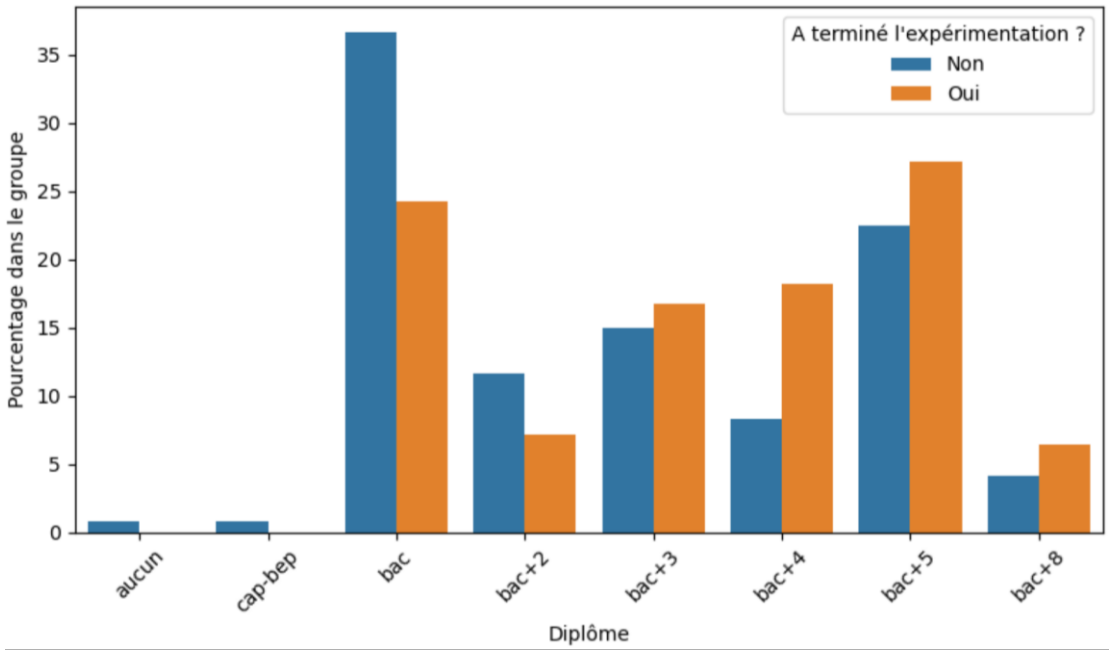

Figure 36 : Répartition des diplômes selon l'abandon du test ou la persévérance du participant

La moyenne globale de crédibilité donnée par les participants qui ont abandonné l'expérimentation (moyenne = 2,59) est plus élevée que celle des participants qui l'ont terminée (2,41).

En s'intéressant au détail par facteurs, nous pouvons voir que les participants qui ont abandonné l'expérimentation ont fait moins de différence entre les facteurs (voir Figure 37). Nous pouvons supposer que ces participants ont moins lu les textes, ou qu'ils ont perçu une forte redondance qui les a poussés à abandonner l'expérimentation.

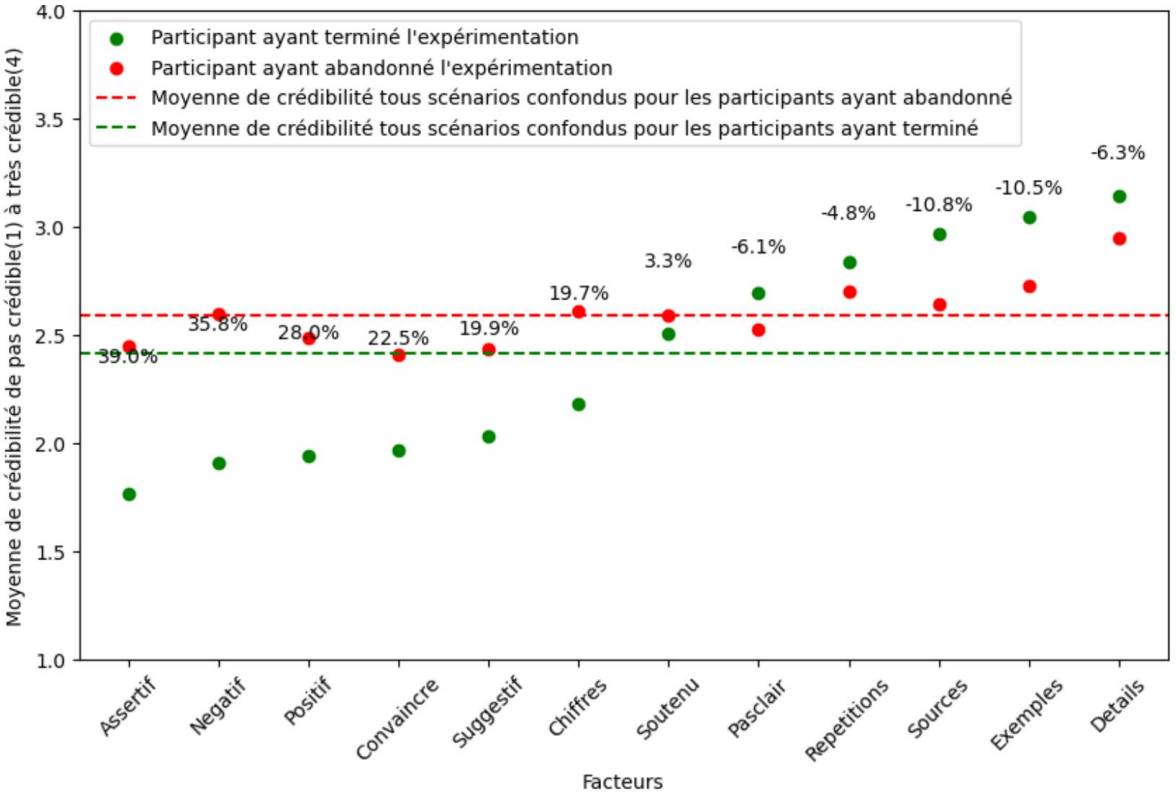

Figure 37 : Moyenne de crédibilité de chaque facteur en fonction de la persévérance du participant

3.7 Discussion

Les résultats de cette étude offrent des perspectives intéressantes sur la perception de la crédibilité des messages textuels, en particulier dans le contexte de la désinformation.

Description des principaux résultats :

Il semble y avoir une corrélation entre le niveau d'études et le niveau de sensibilité à la désinformation (score MIST), les personnes ayant les plus hauts niveaux d'études étant en moyenne légèrement moins sensibles à la désinformation dans la population interrogée. Il semble également exister une différence entre les hommes et les femmes, ces dernières ayant un score MIST moyen et médian plus bas que celui des hommes. Cela pourrait s'expliquer par la présence de plus d'hommes parmi les participants interrogés ayant les plus hauts niveaux d'études (Bac+5 et Bac+8), étant donnée la légère corrélation entre score MIST et niveau d'études.

Les cartes du jeu « Taylor Swift & Selena Gomez » décrivaient de fausses informations tentant de faire croire que les célébrités en question dénonçaient l'aide Occidentale à l'Ukraine. Elles ont été classées comme moins crédibles en moyenne que les cartes des deux autres jeux (décrivant de vraies informations et des faits réels et potentiels), ce qui est cohérent avec le fait qu'elles relayaient une tentative de désinformation évidente pour un lecteur averti, c'est-à-dire une personne disposant d'un niveau de littératie informationnelle et argumentative élevé, habituée à évaluer la fiabilité des sources et la cohérence des raisonnements (Wineburg & McGrew, 2019).

Les facteurs les plus influents, tous scénarios confondus, sont, dans l'ordre : *détails, exemples, sources* et r*épétitions*. Reflétant les résultats de la phase A où ils étaient particulièrement confondus dans le cadre

du scénario 1, *exemples* et *détails* sont très proches en crédibilité pour le même scénario lors de la phase B.

Les facteurs les moins crédibles, tous scénarios confondus, incluent la présence d'*émotions*, d'une intention de *convaincre* et l'*assertivité*.

Les variations observées entre les scénarios peuvent s'expliquer par la manière dont les facteurs ont été implémentés et par la crédibilité intrinsèque des scénarios de base. Cependant, les facteurs *répétitions*, *sources*, *exemples* et *détails* restent les quatre facteurs apportant le plus de crédibilité, même dans le scénario de désinformation pure. Le facteur *pas clair* est particulièrement influent dans le scénario 2 (Algorithmes).

Nous observons quelques différences entre les groupes de participants (Tableau 14) :

- Les personnes ayant un niveau d'études plus élevé accordent plus de confiance à la présence de *sources*, tandis que celles ayant un niveau d'études plus faible se fient davantage que les autres à des messages *pas clairs*.
- Les participants ayant un score MIST faible accordent plus de crédibilité à la présence de *chiffres*, d'*assertivité*, d'*émotions positives* et à des messages *pas clairs*, tandis que ceux ayant un score MIST élevé privilégient à nouveau la présence de *sources*.
- Aucune différence significative n'a été observée entre les différents groupes de rattachement (étudiants, collaborateurs THALES, Ministère des Armées, etc.).

Tableau 14 : Différence d'importance des facteurs de crédibilité entre les groupes de participants

Population	**Tous**	Niveau d'études >Bac+3	Niveau d'études <=Bac+3	Score MIST >15/20	Score MIST <=15/20
Facteurs + crédibles	Détails Exemples Sources Répétitions	Sources	Message pas clair	Sources	Message pas clair
Facteurs - crédibles	Intention de convaincre Émotions Assertivité			Chiffres Assertivité Émotions positives	

Interprétation :

La consigne « d'évaluer la crédibilité d'un message » fait directement appel à l'esprit critique des participants. Or, les personnes ayant un niveau d'études ou un score MIST élevé ont souvent appris à vérifier la présence de sources, permettant de rendre l'information vérifiable et plus crédible. Cela correspond au modèle de traitement de l'information proposé par Petty et Cacioppo (1984) dans le *Elaboration Likelihood Model*, selon lequel la crédibilité d'un message dépend de la voie de traitement empruntée : centrale, fondée sur l'analyse des arguments (présence de sources, cohérence, exemples), ou périphérique, fondée sur des indices plus superficiels (style, émotions, répétitions). La présence de détails et d'exemples peut également donner une impression de crédibilité plus importante : peu de détails ou d'exemples peuvent indiquer un manque d'arguments, alors que plus de détails permettent au lecteur de mieux forger sa propre opinion.

La présence de répétitions pourrait introduire un biais cognitif, connu sous le nom d'effet de vérité, où une information répétée semble plus vraie. Béna et al. (2019) ont montré que cet effet est particulièrement puissant pour des textes courts (comme ceux considérés dans notre étude) et à court terme. Il peut perdre son efficacité dans le temps.

La présence d'émotions, d'un langage assertif et d'une intention de convaincre apparente sont plus faciles à détecter comme une tentative de manipulation par un lecteur averti, étant connues comme des techniques d'influence. Nous pouvons supposer que cela a influé sur leur classement comme des facteurs moins crédibles.

Implications des résultats :

Nos résultats fournissent une première hiérarchisation des facteurs les plus crédibles, présentant donc le potentiel d'influence le plus élevé. Les facteurs *détails*, *exemples*, *sources* et *répétitions* améliorent la crédibilité tant des informations réelles que fausses.

Lors de l'étape de validation du matériel (phase A), certains participants ont exprimé un intérêt particulier pour la tâche de tri de cartes qui leur a été confiée en indiquant qu'associer chaque carte à sa définition leur permettait de se rendre compte des tentatives de manipulation. Ils reconnaissaient des formulations déjà rencontrées. Nous pourrions imaginer en faire un outil de formation et de prévention de certaines techniques de manipulation, sur le même principe que le plugin[6] du projet M82 utilisant la matrice DIMA (voir *1.4.3 DIMA*) pour identifier les manipulations cognitives en ligne.

Utilisation des résultats :

Imaginons un texte de désinformation, inspiré du réseau de faux comptes « COP 29 » qui visait à promouvoir l'Azerbaïdjan dans le cadre de son accueil du sommet sur le climat de la COP29 :

> *L'Azerbaïdjan est un exemple à suivre pour l'Europe en matière de politiques écologiques. Ce pays a su allier développement économique et respect de l'environnement, contrairement à la France.*

Une version jugée plus crédible, incorporant quelques détails et exemples, des sources et des répétitions, pourrait être :

> *L'Azerbaïdjan est un exemple à suivre pour l'Europe en matière de politiques écologiques, comme l'indiquent plusieurs rapports internationaux sur l'innovation verte. Ce pays a su allier développement économique et respect de l'environnement, contrairement à la France qui échoue à mettre en place des politiques efficaces pour la transition énergétique, comme en témoignent la lenteur de la fermeture de ses centrales à charbon et l'inefficacité des mesures prises contre la pollution de l'air.*

Limites de l'étude :

Notre échantillon, jeune et composé principalement de personnes diplômées et d'étudiants, présente un biais éducatif. Cette population est souvent plus critique face à la désinformation. Cela limite la généralisation des résultats à des populations moins éduquées. Cette répartition est cohérente avec les canaux utilisés pour diffuser l'étude (entreprise, écoles, réseaux personnels et professionnels), mais aussi avec notre objectif : en effet, ces travaux de thèse s'intéressent principalement aux décideurs et aux

[6] https://m82-project.org/articles/dima/plugin_dima_chrome/

leaders, qui ont souvent un niveau d'études ou d'expérience élevé. Dans les deux cas, il s'agit d'une population qui a tendance à être moins crédule à la désinformation, qui est le sujet de cette étude. De plus, la consigne d'évaluer la crédibilité place les participants dans un contexte différent de leur consommation quotidienne d'informations en faisant directement appel à leur esprit critique. Cette approche introduit une limitation quant à l'applicabilité des résultats à des contextes de consommation passive d'informations. Cependant, cette méthode se rapproche d'une situation d'analyse de l'information, similaire à celle rencontrée dans des contextes de prise de décision informée.

Par ailleurs, nous observons une légère sensibilité au scénario de base et donc à l'implémentation des facteurs dans la rédaction des textes. Cet effet semble limité dans nos résultats, mais les facteurs textuels considérés n'étant pas totalement objectifs, ils pourraient être implémentés différemment par un autre panel d'experts et engendrer des résultats différents. De plus, les sujets et formats des textes sélectionnés sont relativement similaires. Nous ne pouvons donc pas généraliser les résultats à n'importe quel contexte. Par exemple, un texte traitant de l'utilisation des fonds publics pourrait gagner en crédibilité avec l'inclusion de données chiffrées précises.

L'une des principales limitations de cette étude réside dans l'absence de croisement des facteurs, ce qui aurait permis d'explorer les effets d'interaction possibles. Par exemple, un message combinant la présence de sources et un appel discret aux émotions pourrait attirer davantage l'attention du lecteur et être perçu comme plus crédible qu'un message se contentant de citer ses sources. Cette absence d'interaction limite notre compréhension des nuances potentielles dans la manière dont les lecteurs évaluent la crédibilité des messages.

Cette étude ne traite pas de l'évolution des facteurs de crédibilité dans le temps. Bien qu'ils semblent avoir un potentiel d'influence significatif à court terme, il serait pertinent d'explorer leur impact à moyen et long terme. En l'état actuel, nous ne pouvons pas conclure sur la manière dont ces facteurs continuent d'influencer le lecteur après l'étude.

Ainsi, plusieurs pistes ont été identifiées pour des recherches futures :

- Explorer les interactions entre les facteurs identifiés comme les plus influents.
- Étudier des populations moins diplômées ou plus sensibles à la désinformation.
- Tester différents formats de textes et contextes de consommation d'informations.
- Réintroduire les facteurs écartés concernant l'auteur du message et le média utilisé, ainsi que des facteurs concernant d'autres formats d'information : vocal, vidéo, image….
- Étudier l'évolution de la crédibilité des facteurs dans le temps : après l'étude, lesquels sont le mieux remémorés ?

3.8 Conclusion de l'expérimentation 1

Notre étude recense de nombreux facteurs influençant la crédibilité de messages textuels et propose une tentative de classement. Elle souligne l'importance de la présence de détails, d'exemples, de sources et de répétitions dans la perception de la fiabilité d'un texte. À l'inverse, l'expression d'émotions, l'assertivité ou une intention manifeste de convaincre tendent à diminuer cette crédibilité.

Le niveau d'études semble légèrement corrélé à la capacité de détection de la désinformation, ainsi qu'à la valeur accordée à la présence de sources. Ces résultats contribuent à comprendre comment la désinformation peut être détectée ou au contraire, renforcée. Ils rappellent également la nécessité d'une

formation à la vérification de sources et à la détection de la désinformation. L'importance de l'éducation à l'esprit critique dès l'enfance est formalisée depuis au moins un siècle (Dewey, 1910) et soulignée par de nombreux auteurs. Elle est confirmée par plusieurs études empiriques : une stimulation adéquate de l'esprit critique réduit l'attrait de certaines propositions trompeuses, comme les théories du complot (Swami et al., 2014 ; Gervais & Norenzayan, 2012). Elle constitue un enjeu central, dans un contexte où la surabondance d'informations et la viralité des réseaux sociaux exposent chacun, dès le plus jeune âge, à des contenus non vérifiés et souvent manipulatoires.

Chapitre 4 - Élaboration d'un système pour la détection en temps réel et l'influence des stratégies – exemple du jeu de société Galèrapagos

4.1 Introduction

Chaque décision que nous prenons est conditionnée par notre conscience de la situation et par nos connaissances préalables. Selon Endsley (1995), la conscience de situation est *« la perception des éléments de l'environnement dans un volume de temps et d'espace, la compréhension de leur signification et la projection de leur statut dans le futur proche »*. Elle se décline en 3 niveaux : perception, compréhension, anticipation.

Dans un cadre de gestion de crise, il est nécessaire d'avoir une conscience de la situation juste et adaptée afin d'anticiper l'évolution des événements et d'adapter la prise de décision (Prébot, 2020), ainsi que suivre des processus qui protègent la décision de diverses failles cognitives ou des erreurs possibles. Les erreurs cognitives découlent souvent des heuristiques, que nous utilisons pour prendre des décisions plus efficacement, mais qui peuvent parfois mener à des erreurs de jugement. Ces dernières peuvent survenir à tout niveau de la chaîne de décision, de la prise d'information à la prise de décision, en passant par la compréhension de la situation (Tversky & Kahneman, 1981 ; Thibodeau & Boroditsky, 2011).

Dans cette étude, nous nous intéressons à l'étape de compréhension de la situation, notamment la détection des premiers indices ou signaux faibles signalant l'apparition d'un changement. Nous étudions la possibilité de s'appuyer sur une détection précoce pour faire évoluer la situation à notre avantage, en utilisant des éléments de la cognition de la personne cible afin d'influencer sa décision.

Dans ce cadre, nous avons étudié la possibilité de détecter la stratégie d'un joueur de Galèrapagos en temps-réel en utilisant le principe du logiciel ANTICIPE développé par THALES pour la planification militaire. Cet outil s'appuie sur une structure d'arbre de détection d'informations critiques à trois niveaux pour améliorer la conscience de situation d'un commandeur. Le niveau le plus bas représente les signaux faibles, ou *« cues »*, qui sont les premiers à laisser entrevoir un changement de situation ; chaque cue a un poids associé. Le niveau intermédiaire contient les *« triggers »*, chacun rassemblant plusieurs cues et étant déclenché lorsque plusieurs cues sont activées, ou une cue avec un poids important (information critique). Le niveau le plus haut de l'arbre contient les informations critiques nécessaires pour la prise de décision, soit les CCIR (Commander Critical Information Requirements) tels que définis par l'OTAN[7]. La personne qui utilise le système peut alors identifier rapidement les problèmes ou tensions détectés par le système sur cet ensemble de points de surveillance critiques. Par exemple (Figure 38), si un acteur X prétend que les données des clients d'un acteur Y ont fuité (*cue 1*) et si X est suspecté d'une tentative de phishing chez Y (*cue 2*), on peut en déduire que X désinforme au sujet de Y (*trigger 1*) et tente de lui voler des informations (*trigger 2*) : l'acteur X agit donc activement contre Y (*CCIR* associé). L'agent qui surveille les actions de X via l'arbre de détection d'informations critiques va donc pouvoir avertir l'acteur Y ou mettre en place des contre-mesures pour le protéger.

[7] https://www.jcs.mil/Portals/36/Documents/Doctrine/fp/ccir_fp4th_ed.pdf?ver=2020-01-13-083331-097

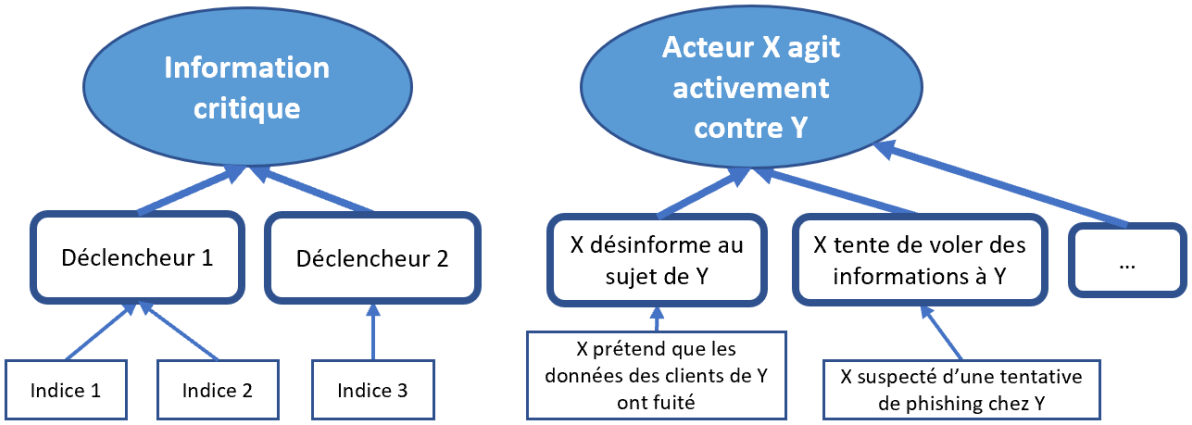

Figure 38 : Exemple d'arbre de détection d'informations critiques de type ANTICIPE

Pour cette étude, notre objectif est d'étudier expérimentalement la possibilité et la pertinence d'adapter ce système au contexte de la guerre cognitive. Comme preuve de concept, il s'agit de tenter de détecter les stratégies et leur influence au cours d'un jeu de société semi-collaboratif.

Nous cherchons ainsi à répondre aux questions de recherche suivantes :

- La méthodologie d'arbre de détection d'informations critiques utilisée par le système ANTICIPE permet-elle d'évaluer la stratégie des joueurs en temps-réel ?
- Peut-on mener une action qui influence les décisions prises par un joueur pendant le jeu (Galèrapagos) ou l'issue de la partie, sans qu'il s'en aperçoive et en s'appuyant sur des éléments de sa cognition et sur les signaux faibles permettant de déterminer sa stratégie ?

4.2 Méthode expérimentale

4.2.1 Choix du support de l'expérimentation

Pour répondre à nos questions de recherche, nous avons initialement identifié deux mises en situation possibles : soit un wargame, soit un jeu de société de stratégie disponible dans le grand public. Un wargame aurait été plus proche de la réalité du terrain et aurait permis de mieux matérialiser les applications possibles du système étudié. Cependant, il est plus long à mettre en place et il peut être plus difficile de trouver des participants à l'expérimentation qui soient suffisamment formés au jeu pour mettre en place des stratégies stables. Nous avons donc fait le choix de nous tourner vers un jeu de société disponible au grand public, qui nous permettait de mettre en place une expérimentation plus rapidement et plus accessible pour des participants non spécialistes.

Nous avons choisi un jeu de stratégie semi-collaboratif nommé Galèrapagos. Ce jeu permet à chaque participant de s'appuyer sur différents types de stratégies et les règles sont accessibles, ce qui nous permet d'obtenir facilement un panel de joueurs satisfaisant. De plus, nous disposions, au sein du laboratoire, d'experts du Galèrapagos, ce qui nous permettait de développer plus facilement l'arbre de détection d'informations critiques nécessaire au système de type ANTICIPE.

➢ Point-clé : principe du jeu Galèrapagos

Galèrapagos est un jeu semi-coopératif dans lequel des naufragés doivent survivre sur une île et construire un radeau pour s'échapper avant que les ressources ne s'épuisent. À chaque tour, chaque joueur choisit une action : pêcher pour obtenir de la nourriture, chercher de l'eau, ramasser du bois pour fabriquer le radeau. L'eau et la nourriture sont consommées totalement ou partiellement à chaque tour. Les joueurs disposent de cartes qui peuvent être utiles pour la communauté (ressources) ou à usage plus individualiste (par exemple, revolver pour éliminer un joueur de son choix).

Règles initiales du jeu Galèrapagos :

Selon le principe de Galèrapagos (Figure 39), des naufragés sur une île déserte doivent construire un radeau pour s'échapper avant l'arrivée d'une tempête. Ils ont des ressources d'eau et de nourriture limitées. Ils disposent chacun de 4 cartes au début de la partie. À tour de rôle, les joueurs doivent choisir une action parmi les suivantes :

- pêcher : le joueur qui choisit cette action a des chances de rapporter 1 à 3 poissons (ou jusqu'à 5 s'il a la carte « canne à pêche »), chaque poisson correspondant à une ration de nourriture pour une personne ;
- recueillir de l'eau : le joueur qui choisit cette action récolte autant de ressources d'eau qu'indiqué sur la carte « météo » du tour, c'est-à-dire 0 à 3 (ou le double s'il a la carte « gourde ») ;
- chercher du bois : le joueur qui choisit cette action récolte par défaut au moins un morceau de bois (ou deux s'il a la carte « hache »), mais peut choisir de piocher une ou plusieurs boules dans une poche contenant 5 boules blanches et 1 boule noire. S'il ne pioche que des boules blanches, il ramènera autant de morceaux de bois supplémentaires, s'il pioche une boule noire, il est « mordu par le serpent » : il perd alors tous ses morceaux de bois supplémentaires et tombe malade. Il ne pourra ni utiliser ses cartes, ni effectuer d'actions jusqu'à la fin du prochain tour ;
- piocher une carte : ces cartes peuvent apporter des ressources supplémentaires ou permettre des actions défensives, offensives, pragmatiques ou collaboratives (cf. Tableau 19 – p. 128). Certaines cartes sont permanentes (comme la « canne à pêche », la « gourde » ou la « hache ») : leur effet est valable pour leur propriétaire jusqu'à la fin de la partie ou jusqu'à son élimination.

Le choix des actions individuelles peut être débattu entre les joueurs. Chaque joueur consomme une ressource d'eau et une ressource de nourriture à la fin de chaque tour. Il faut 6 morceaux de bois (un radeau) et 2 ressources d'eau et de nourriture par personne pour quitter l'île à l'arrivée de la carte « tempête » comme météo du tour. Celle-ci survient aléatoirement à partir du tour 7 (voir les règles détaillées sur le site de Galèrapagos[8]).

[8] https://www.gigamic.com/index.php?controller=attachment&id_attachment=167

Figure 39 : Plateau de jeu de Galèrapagos

Il est nécessaire de faire équipe pour survivre, mais également d'être prêt à sacrifier des équipiers en cas de pénurie (manque d'eau, de nourriture ou pas suffisamment de radeaux à l'arrivée de la tempête) : une élimination peut alors se faire par vote (en cas de pénurie) ou grâce aux cartes « revolver » et « cartouche » si un joueur parvient à les rassembler. Ce jeu fait donc intervenir des mécaniques d'alliances, de trahison et différents profils de joueurs. Différentes stratégies peuvent émerger : collaborative, individualiste, pragmatique ou irrationnelle (déterminées au cours de la phase A de l'expérimentation – voir paragraphe *4.2.2.1 Phase A : Croisement de regards d'experts*).

Adaptation des règles du jeu Galèrapagos à notre expérimentation :

Nous avons adapté le jeu en introduisant une pioche systématique de cartes, ce qui donne plus d'indices sur la stratégie du joueur étant donné que celles-ci peuvent être utiles pour une stratégie individualiste (revolver, cartouche, cartes de protection individuelles), collaborative (ressusciter un autre joueur, apporter des ressources au groupe) ou pragmatique (transformer de l'eau en nourriture ou l'inverse, ne pas faire d'action à ce tour mais en faire deux plus tard…) (Morelle-Gerritsen et al., 2025). Ainsi, chaque joueur consulte une carte sans la montrer aux autres avant de choisir son action (récolte d'eau, de nourriture ou de bois) ; s'il fait le choix de la conserver, il renonce à son action ; s'il la défausse, il pourra choisir une action.

4.2.2 Protocole expérimental

Pour répondre à nos questions de recherche, nous avons conçu un protocole expérimental en trois phases.

Les deux premières phases consistent à préparer le matériel d'expérimentation et la troisième cherche à répondre aux questions de recherche (Figure 40). La Phase A - Croisement de regards d'experts et jeu de parties informelles avec des rôles-types, a pour objectif d'identifier les différents types de stratégies possibles et quelles actions traduisent quelles stratégies ; cette phase permet de construire une ébauche de l'arbre de détection d'informations critiques. La Phase B - Jeu de parties sur plateau avec 12 participants, permet d'enrichir et de valider l'arbre de décision conçu lors de la phase A. La Phase C - Jeu sur version numérique avec des participants isolés, a pour but de détecter la stratégie du joueur en temps réel grâce à l'arbre de décision ; puis de tenter de l'influencer vers la stratégie opposée à la sienne.

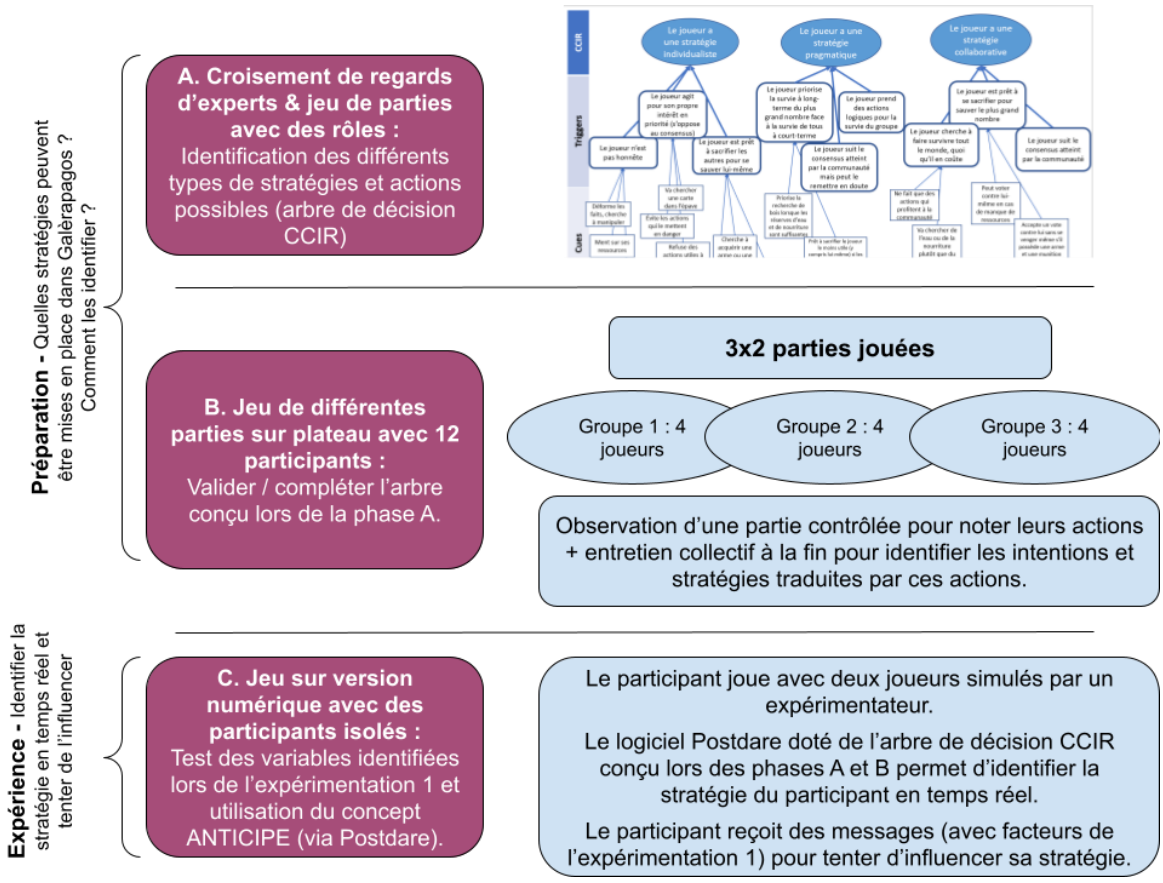

Figure 40 : Représentation schématique du protocole expérimental en 3 phases

4.2.2.1 Phase A : Croisement de regards d'experts (développement de l'arbre de détection d'informations critiques)

Le matériel utilisé au cours de la phase A consiste en :

- 1 jeu de Galèrapagos (plateau)
- 1 caméra (Handy Video Recorder Q2n-4K ZOOM)

5 experts et 3 non experts du jeu de Galèrapagos ont participé à cette phase. L'objectif est d'identifier les actions possibles dans le jeu, puis de tenter d'établir un premier lien avec les stratégies dont elles relèvent.

Pour ce faire, nous avons mis en place un *focus group*[9] (Morgan, 1997) de 5 experts du jeu de Galèrapagos qui a participé à plusieurs activités visant d'abord à identifier les stratégies possibles dans le jeu à partir de leur propre expérience, puis à lister les actions possibles en confrontant leurs visions sur le jeu. Ce groupe d'experts a identifié quatre stratégies possibles dans le jeu de Galèrapagos : individualiste (privilégier sa propre survie), pragmatique (faire survivre le groupe si possible, sinon soi-même), collaboratif (privilégier la survie du groupe complet) ou un profil irrationnel (comportement illogique avec des actions qui ne favorisent ni sa propre survie, ni celle du groupe).

[9] Focus group : groupe réuni dans le cadre d'une discussion structurée pour examiner un sujet et en dégager un consensus.

Le focus group a ensuite joué plusieurs parties, certaines avec des joueurs non experts. Chaque participant avait pour rôle de jouer une stratégie en particulier parmi celles qui ont été identifiées. Ainsi, nous avons organisé des parties avec uniquement des joueurs ayant un rôle collaboratif, uniquement individualiste ou pragmatique, et en croisant les différentes stratégies et niveaux d'expertise possibles. Cela nous a permis d'identifier un grand nombre d'actions possibles et les liens entre actions et stratégies.

4.2.2.2 Phase B : Parties sur plateau avec 12 participants (enrichissement de l'arbre de détection d'informations critiques)

Le matériel utilisé au cours de la phase B consiste en :

- 2 jeux de Galèrapagos (plateau)
- Questionnaires en ligne (Google Forms) : questionnaires GDMS (style de prise de décision), Big Five (personnalité) et MCQ-30 (métacognitif) (présentés ci-après)
- 1 caméra (Handy Video Recorder Q2n-4K ZOOM)

12 joueurs de Galèrapagos expérimentés (4 pour chaque passation de l'expérimentation) ont participé à cette phase.

2 expérimentateurs ont été impliqués dans la mise en œuvre du protocole : un expérimentateur principal dirige les participants sur les différentes tâches à effectuer, réalise l'observation du déroulement de la partie et conduit l'entretien ; un deuxième expérimentateur assiste dans la mise en place des plateaux de jeu, gère la caméra et observe le comportement non-verbal et les interactions des joueurs au cours des parties.

L'objectif de cette phase B était de confronter et compléter les résultats de la phase A. Dans ce but, nous avons fait jouer trois groupes de quatre joueurs expérimentés. Pour parfaire leur connaissance du jeu, nous avons organisé au préalable un tournoi de Galèrapagos avec l'obligation pour chaque participant de jouer au moins quatre parties au cours du tournoi. Ainsi, les participants étaient très familiarisés avec le jeu et avaient eu l'occasion de développer leurs propres stratégies pour maximiser leurs gains au cours du tournoi.

Le protocole de la phase B est décrit dans le Tableau 15.

Les prérequis pour les participants sont d'être majeur, de comprendre et parler le français couramment, et d'avoir joué au moins quatre parties de Galèrapagos récemment, au cours du tournoi que nous avons organisé au préalable.

Nous commençons par donner aux participants les consignes de l'expérimentation, leur en décrire le déroulement puis leur faire signer un formulaire faisant état de leur droit de retrait, des mentions RGPD et recueillant leur consentement explicite.

Tableau 15 : Protocole détaillé de la phase B en 10 étapes

Étape	Données à récolter	Durée
1. Short French Metacognition questionnaire	Structures des croyances métacognitives de régulation utilisées par les joueurs, telles que la planification, le suivi, la régulation des erreurs et la réévaluation des stratégies	3 min
2. Questionnaire de personnalité : Big Five inventory	Mesure des cinq principaux traits de personnalité (extraversion, agréabilité, conscienciosité, ouverture à l'expérience, neuroticisme)	5 min
3. Questionnaire GDMS	Mesure les styles décisionnels : rationnel, intuitif, dépendant, évitant et spontané	5 min
4. Partie 1 – entraînement (rappel des règles ; arrivée de la tempête au bout de 4 tours pour écourter)	Aucune	15 min
5. Questionnaire main initiale idéale	4 cartes idéales pour démarrer une partie + explication	2 min
6. Partie 2 : contrôle des mains initiales et de la pioche pour permettre différents choix de stratégies	Prise vidéo Observation : données quantitatives du jeu, verbatims et communications verbales, dynamique de l'affectivité groupale, activité corporelles et comportementale	30 min
7. Questionnaire de ressenti général du joueur	Vision globale du participant sur la partie et sa propre stratégie	3 min
8. Questionnaire démographique	Âge, niveau d'études, genre, nationalité, niveau d'expertise au Galèrapagos	2 min
9. Questionnaire MIST	Sensibilité de chaque joueur à la mésinformation (score total), scores de faux positifs et faux négatifs	5 min
10. Entretien collectif	Vision de chaque participant sur sa propre stratégie et celle des autres	60 min
Total		**2h10**

Étapes 1 à 3 : Questionnaires de personnalité, métacognition et style de prise de décision.

Ces questionnaires doivent permettre d'évaluer les profils psychologiques et métacognitifs des joueurs dans leur vie quotidienne. Ils sont donc donnés au début de l'expérimentation, en amont de la partie de Galèrapagos afin d'éviter d'observer un effet du déroulement des parties sur les réponses données. Ces questionnaires sont réalisés en présence d'un examinateur.

Pour obtenir une large description du profil du joueur, nous choisissons d'évaluer son style métacognitif et psychologique. Différents questionnaires disponibles sont comparés dans le Tableau 16 et le Tableau

17 respectivement. Les questionnaires retenus pour l'expérience sont colorés en bleu dans les tableaux ci-dessous.

Tableau 16 : Comparaison des questionnaires métacognitifs étudiés

Questionnaires métacognitifs	**Metacognitive Awareness Inventory (MAI)**	**Cognitive Control and Flexibility Questionnaire (CCFQ)**	**Short French Metacognition questionnaire (MCQ-30)**
Références	Schraw & Dennison, 1994	Gabrys et al., 2018	Baptista et al., 2014 ; Dethier et al., 2017
Avantages	- Diversité des données métacognitives mesurées - Validité scientifique	- Mesure la régulation mentale et la flexibilité cognitive et émotionnelle	- Version française validée - Durée (30 questions) - Spécifique aux croyances métacognitives
Inconvénients	- Pas de version française validée - Durée (52 questions)	- Pas de version française validée - Spécifique au contrôle des pensées et réactions émotionnelles face au stress	- Principalement étudié dans un cadre clinique

Le Tableau 16 permet de constater que :

- Le *« Metacognitive Awarness Inventory »* (MAI – Schraw & Dennison, 1994) n'est disponible qu'en anglais sans version validée en français, ce qui pose un problème pour sa mise en œuvre dans le cadre de notre expérimentation.
- Le *« Cognitive Control and Flexibility Questionnaire »* (CCFQ – Gabrys et al., 2018) est spécifique au contrôle des pensées et réactions émotionnelles face au stress. Il pourrait tout de même présenter un intérêt pour notre expérimentation afin d'étudier le lien entre la flexibilité cognitive et l'adaptation de la stratégie à de nouvelles situations de jeu, mais il ne dispose pas non plus de version validée en français.
- Le questionnaire *« Short French Metacognition questionnaire »* (MCQ-30 – Baptista et al., 2014 ; Dethier et al., 2017), quant à lui, évalue la perception des participants en ce qui concerne leur propre manière de penser. Il mesure en particulier cinq aspects :
    - Cognitive confidence (confiance ou manque de confiance cognitive) ;
    - Positive beliefs about worry (croyances positives à propos de l'inquiétude) ;
    - Cognitive self-consciousness (conscience cognitive de soi) ;
    - Negative beliefs about uncontrollability and danger (croyances négatives concernant l'incontrôlabilité et le danger) ;
    - Beliefs related to Superstition, Punishment and Responsibility (croyances concernant la superstition, punition et responsabilité).

  Certains de ces aspects peuvent être d'intérêt pour notre expérimentation : en effet, la confiance cognitive ou encore le besoin de contrôle cognitif pourraient être liés à une plus grande vulnérabilité ou résistance à l'influence.

Nous sélectionnons donc le questionnaire MCQ-30 dans le cadre de cette étude, afin d'évaluer les croyances métacognitives des participants et leur éventuel lien avec les stratégies adoptées au cours du jeu.

Tableau 17 : Comparaison des questionnaires de profil psychologique

Questionnaires psychologiques	Big five Inventory (BFI)	Échelle d'Individualisme-Collectivisme (INDCOL)	General Decision-Making Style (GDMS)
Références	John et al., 1991 ; Plaisant et al., 2010	Hui, 1988	Scott & Bruce, 1995 ; Girard et al., 2016
Avantages	- Mesure complète de la personnalité - Version française validée - Largement documenté	- Spécifique à l'orientation sociale (coopération vs individualisme)	- Version française validée - Mesure de la prise de décision
Inconvénients	- Durée (45 questions) - Peu spécifique à la décision ou à l'influence - Ne mesure pas directement les dynamiques sociales	- Pas de version française validée	- Spécifique à la prise de décision - Ne mesure pas la posture sociale

Le Tableau 17 permet de constater que :

- Le questionnaire *« Big Five Inventory »* (BFI – John et al., 1991 ; Plaisant et al., 2010) évalue cinq grands traits de personnalité :
    - extraversion, énergie, enthousiasme (orientation sociale et niveau d'énergie) ;
    - agréabilité, altruisme, affection (tendance à être compatissant, coopératif et bienveillant) ;
    - conscience, contrôle, contrainte (niveau d'organisation, discipline et fiabilité) ;
    - émotions négatives, névrosisme, nervosité (stabilité émotionnelle, un score élevé sur ce trait peut révéler un caractère anxieux, triste ou colérique) ;
    - ouverture, originalité, ouverture d'esprit (propension à être curieux, imaginatif, créatif et ouvert à de nouvelles expériences).

  Le questionnaire Big Five est le plus générique des trois questionnaires comparés, permettant d'évaluer des traits de personnalité variés et généraux. Nous choisissons donc de conserver ce questionnaire afin d'évaluer si certains traits peuvent contribuer à prédire les tendances stratégiques (voir paragraphe *4.3.3.6 Révision du modèle de prédiction de la stratégie à partir du profil psychologique, métacognitif et décisionnaire*).

- Quant à l'échelle *« Individualisme-Collectivisme »* (INDCOL – Hui, 1988), mesurant spécifiquement les tendances collectivistes ou individualistes, elle aurait pu être particulièrement intéressante pour l'ébauche de modèle de prédiction des stratégies. Cependant, ce modèle ne constituait pas un objectif prioritaire de notre expérimentation, et l'utilisation de cette échelle est compliquée par l'absence de traduction française validée (à notre connaissance). Ce questionnaire a donc été écarté mais reste un outil intéressant pour de futures études.

- Le questionnaire *« General Decision-Making Style »* (GDMS – Scott & Bruce, 1995 ; Girard et al., 2016) présente un intérêt dans l'évaluation des motivations à la prise de décision et pourra donc être particulièrement utile pour notre étude, qui s'intéresse à l'influence des décisions stratégiques dans le jeu. Le questionnaire GDMS évalue le ou les styles de prise de décision de la personne parmi cinq différents styles :
    - rationnel (suit une approche logique, structurée et basée sur les faits) ;
    - intuitif (décide en fonction de son instinct, ses impressions ou son expérience personnelle) ;
    - dépendant (recherche l'avis et le soutien des autres) ;
    - évitant (tend à reporter ou éviter la prise de décision) ;

- spontané (décide rapidement, parfois de manière impulsive et en fonction des circonstances immédiates).

Nous choisissons de le conserver afin de compléter les profils psychologiques des joueurs mais aussi en prévision d'une modification de leur décision au cours de la phase C de l'expérimentation.

Nous sélectionnons donc les questionnaires Big Five et GDMS dans le cadre de cette étude, afin d'évaluer la personnalité et le style de prise de décision du participant.

Étape 5 : Questionnaire de la main initiale idéale.

Le questionnaire de la main idéale consiste en deux questions :

1. *Pour débuter une partie de Galèrapagos, ma main initiale idéale serait : (Sélectionner 4 cartes)*
2. *Pour quelle(s) raison(s) avez-vous choisi ces cartes en particulier ?*

Ce questionnaire vise à identifier les préférences stratégiques des joueurs sans influence de la partie ni des interactions avec les autres joueurs ou de la main qui leur a été distribuée. Il intervient entre la partie d'entraînement et la partie enregistrée pour que les joueurs aient les cartes et le jeu en tête. Le set de cartes proposé lors de la première question contient toutes les cartes du jeu ayant un intérêt stratégique potentiel, afin d'éviter toute influence des choix des expérimentateurs.

Étapes 4 et 6 : Les joueurs ont pour consigne de *« quitter l'île soi »* (donc seul), formulation volontairement orientée vers des stratégies plus individualistes car nous avons observé des comportements très majoritairement collaboratifs au cours du tournoi préalable. Afin de permettre aux expérimentateurs de contrôler la main des joueurs à l'étape 6, toutes les parties sont jouées avec les mêmes mécanismes de distribution des cartes et de jeu. Les variables tirées au sort (poisson, bois) sont énoncées à voix haute par le joueur lui-même ou par un expérimentateur pour l'enregistrement vidéo.

La distribution des cartes respecte les procédés suivants :

- 4 cartes par joueur sont distribuées en début de partie (règles initiales du Galèrapagos).
- Chaque joueur possède une pioche personnelle.
- Lorsque c'est son tour, le joueur regarde la première carte de sa pioche et a le choix entre la garder ou la défausser (règle introduite pour cette expérimentation) pour effectuer une action parmi pêcher, récolter de l'eau ou récolter du bois.
- Si lors de son tour le joueur est malade et décide de garder la carte piochée, il reste malade pour le tour suivant.

Pour l'étape 6, les mains initiales et pioches des joueurs sont contrôlées et disposées suivant la grille présentée en Tableau 18.

Grâce à cette distribution équilibrée et donnant aux joueurs des cartes variées (cartes collaboratives, ressources, pragmatiques, individualistes et de protection), nous espérons donner accès à un maximum de stratégies possibles à chaque tour pour tous les joueurs afin de les laisser exprimer leur intention de prédilection.

Tableau 18 : Distribution des cartes des 4 joueurs pour la partie observée de la phase B

	Joueur 1	Joueur 2	Joueur 3	Joueur 4
Set de cartes de départ (4 cartes)	Revolver	Revolver	Revolver	Revolver
	Plaque de tôle	Plaque de tôle	Plaque de tôle	Plaque de tôle
	Sandwich	Bouteille d'eau	Sandwich	Bouteille d'eau
	Panier garni	Kit BBQ Cannibale	Moulin à légumes	Nouilles chinoises
Distribution tour 1	Planche de bois	Clé de voiture de luxe	Ticket de loterie gagnant	Hache
Distribution tour 2	Cartouche	Cartouche	Conque	Jeu de société Quoridor
Distribution tour 3	Anti-venin	Sandwich	Canne à pêche	Cartouche
Distribution tour 4	Bouteille d'eau	Gourde	Cartouche	Taser
Distribution tour 5	Longue-vue	Allumettes	Magazine minceur	Sandwich
Distribution tour 6	Vieux slip	Somnifères	Bouteille d'eau	Eau croupie
Distribution tour 7	Clé de voiture	Ticket de loterie gagnant	Poisson pourri	Brosse à WC

Dans les deux parties, la météo du tour est déterminée à l'avance afin de contrôler le niveau de difficulté de la partie ainsi que sa durée : en effet, plusieurs tours consécutifs avec une météo de 0 (soit pas de pluie) peuvent mettre les joueurs en grande difficulté car ils ne pourront alors pas récolter d'eau ; la durée est contrôlée approximativement en choisissant quand intervient la carte météo « tempête » qui signale la fin du jeu. Pour l'étape 4 (rappel des règles), la tempête arrive en avance, soit au tour 4 afin d'écourter cette partie. Pour l'étape 6 (partie observée), la tempête arrive au tour 7.

Un expérimentateur est également chargé de remplir la grille de jeu qui permet d'inscrire les actions choisies par les joueurs et de garder ainsi une trace de l'entièreté de la partie et de la réutiliser comme guide et source d'information lors de l'entretien collectif. L'autre expérimentateur remplit une grille d'observation des comportements des joueurs et pourra ainsi compléter poser des questions complémentaires lors de l'entretien collectif.

Étape 7 : Le questionnaire de ressenti général et les suivants sont distribués après la partie observée. En particulier, le questionnaire de ressenti est distribué avant les autres pour permettre sa lecture et sa prise en compte par les expérimentateurs avant le début de l'entretien collectif.

L'objectif du questionnaire de ressenti général du joueur est de recueillir un premier ressenti sur la partie sans influence des avis et discussions avec les autres joueurs. Il est composé des 3 questions suivantes :

1. *Avez-vous ressenti une dynamique de groupe particulière dans cette partie ? (Parmi collective, individualiste, ambiguë et autre/réponse ouverte)*
2. *Avez-vous senti à certains moments que la partie prenait une tournure décisive ? Si oui, pourquoi ? (Réponse ouverte)*
3. *Aviez-vous des objectifs en débutant cette partie ? Si oui, lesquels ? Y a-t-il des moments où vous les avez perdus de vue ? (Réponse ouverte)*

La deuxième question permet de relever les moments importants semblant être décisifs dans la partie selon l'opinion du joueur. La réponse pourra être utilisée afin de cibler ces moments au cours de l'entretien collectif et d'approfondir les causes des actions. La troisième question porte sur les objectifs

individuels des joueurs, ce qui doit permettre de caractériser leur stratégie. Elle est complétée par une question sur le maintien des objectifs qui pourra suggérer des leviers d'influence de la stratégie au cours de la phase C.

Étape 8 : Un questionnaire démographique est intégré de manière traditionnelle au protocole afin de dresser le profil des participants et de récolter des informations utiles pour expliquer ou tempérer les résultats de l'expérimentation. Les données récoltées sont : l'âge, le niveau d'études, le genre, la nationalité, le niveau au jeu de Galèrapagos et la fréquence de participation à des jeux de société.

Étape 9 : Le questionnaire MIST, ou Misinformation Susceptibility Test, est le seul questionnaire mesurant la désinformation que nous avons trouvé dans la littérature au moment du lancement de cette expérimentation (Maertens et al., 2023). Nous l'utilisons en préparation de la phase C, dans le but de connaître les joueurs sensibles à la désinformation et donc plus susceptibles d'être influencés via les messages utilisés. De plus, ce dernier questionnaire vient compléter le profil psychologique du joueur qui doit, grâce aux données récoltées, permettre d'établir un modèle de prédiction de la stratégie.

Étape 10 : L'entretien collectif d'auto-confrontation consiste en le visionnage de la partie jouée, accompagné de questions posées par un expérimentateur sur les choix d'actions, de stratégies et de communication des joueurs. L'expérimentateur guide la conversation entre les participants afin de les laisser confronter leurs opinions et récolter des informations sur les stratégies et dynamiques de jeu. Cet entretien doit durer au maximum 60 minutes ; afin de maîtriser la durée, l'expérimentateur qui conduit l'entretien pourra décider de visionner uniquement les tours pendant lesquels se sont déroulés des événements importants.

Dans cette phase B, nous avons choisi de mener un entretien collectif plutôt que des entretiens d'auto-confrontation individuels pour permettre aux joueurs de confronter leurs opinions sur la partie jouée et sur leurs propres stratégies, tout en prenant garde au biais de désirabilité sociale.

4.2.2.3 Phase C : Détecter la stratégie en temps réel et tenter de l'influencer

Le matériel et les outils utilisés pour la réalisation de la phase C comprend :

- 1 jeu de Galèrapagos au format numérique, appelé Panadarchipel (Figure 41), que nous avons développé. Cette version numérique dispose de toutes les fonctionnalités du jeu de plateau (actions, jouer ou donner des cartes…) et d'un journal de bord qui présente l'état du plateau (ressources disponibles et météo) au début de chaque tour, les actions effectuées et cartes jouées par chaque joueur, et via lequel les joueurs peuvent communiquer. Le participant dispose de messages préparés qu'il est encouragé à utiliser (par exemple, *« J'ai une carte X »*, *« Quelle carte as-tu piochée ? »*, *« Joueur X devrait faire action Y »*, etc.) afin de permettre l'automatisation des *« cues »* dans le logiciel Postdare (voir point suivant). Les ordres de distribution des cartes et des météos sont prédéterminés. Il existe deux interfaces dans Panadarchipel : une interface « participant », utilisée par ce dernier, qui ne voit que ses propres cartes comme un joueur normal, et une interface « expérimentateur », qui contrôle les deux joueurs simulés et peut voir les cartes de tous les participants en simultané.

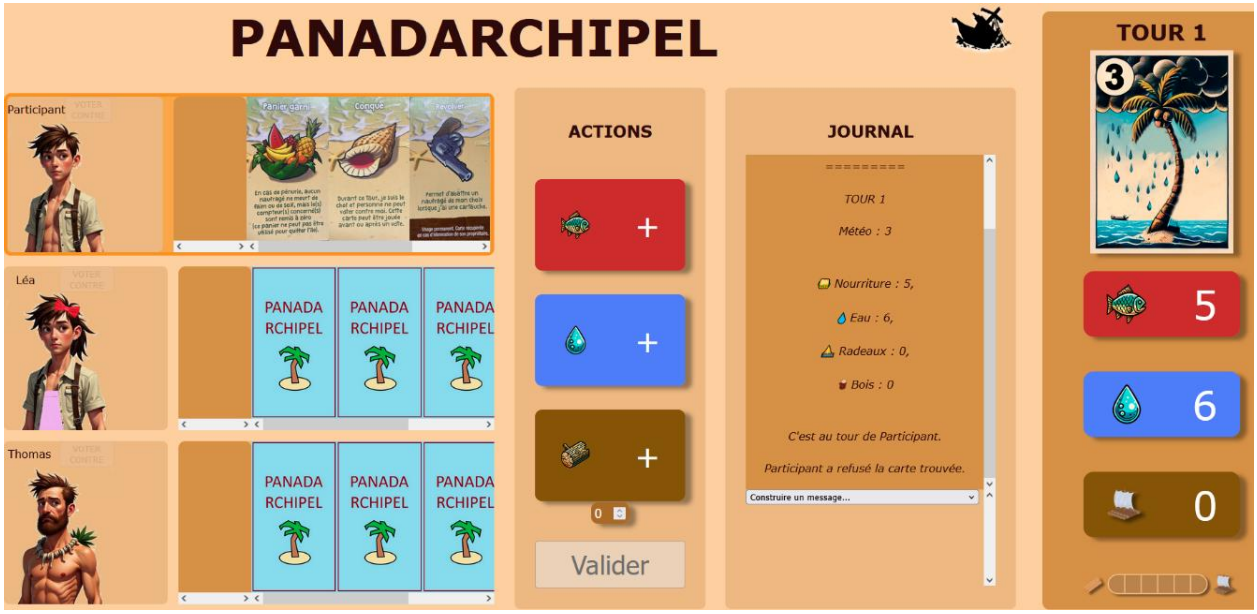

Figure 41 : Interface de Panadarchipel, vue du participant avec de gauche à droite : ses cartes disponibles, ses choix d'actions possibles pour chaque tour, le journal de bord enregistrant les actions et conversations, et l'état du plateau à l'instant T.

- 1 outil implémentant l'arbre de détection d'informations critiques résultant des phases A et B et du principe d'ANTICIPE, appelé Postdare et que nous avons également développé. Il comprend deux interfaces. La première permet de visualiser la totalité des *« cues »* par colonnes (en fonction du thème, comme *« utilisation des cartes »* ou *« débats avec les autres joueurs »*) (Figure 42) et de sélectionner les *« cues »* manuelles. La seconde interface permet de visualiser l'arbre de décision et les pourcentages d'activation de chaque type de stratégie (collaboratif, pragmatique, individualiste ou irrationnel) (Figure 43).

Figure 42 : Interface de Postdare pour l'expérimentateur – affichage des « cues » par catégorie pour sélection manuelle (en rouge) et contrôle de la sélection automatique (en blanc). Les « cues » manuelles apparaissent en premier dans l'affichage de chaque colonne.

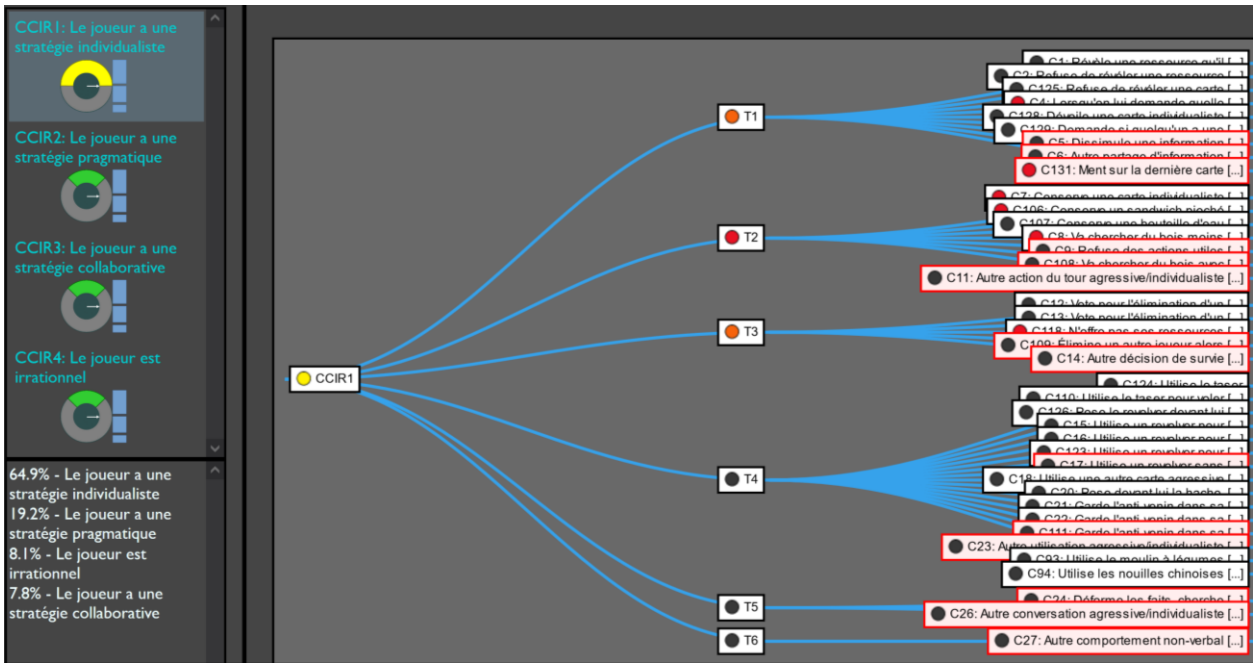

Figure 43 : Interface de Postdare pour l'expérimentateur – affichage de l'arbre de détection d'informations critiques et de la stratégie détectée par le système – exemple d'un joueur individualiste, affichage du CCIR « individualiste » uniquement

- 4 parties préparées (1 partie d'entraînement et 3 parties observées), avec contrôle de la météo et des cartes tirées à chaque tour. Ces parties sont conçues pour que le participant puisse choisir sa propre stratégie, avec notamment des tirages de cartes équilibrés et la création de situations variées et de tensions qui le forceront à faire des choix. Les différentes cartes utilisées sont listées dans le Tableau 19. Les parties préparées sont visibles en *Annexe 2*.

Tableau 19 : Liste des cartes utilisées avec leurs fonctions et leur fréquence d'apparition dans le jeu

Cartes offensives	**Revolver** (rare, carte permanente) **+ cartouches** (courante) : l'association des deux permet d'éliminer un autre joueur. **Somnifères** (rare) : permet de voler 1 carte à chaque autre naufragé. **Longue-vue** (rare) : permet de voir le jeu d'un autre naufragé. **Taser** (rare) : permet de voler une carte à effet permanent posée devant un autre naufragé. Peut aussi avoir un usage pragmatique.
Cartes défensives	**Plaque de tôle** (rare) : permet de se défendre d'un tir de revolver. **Conque** (rare) : permet de se protéger d'un vote.
Cartes ressources	**Panier garni** (rare) : permet de faire manger et/ou boire tout le monde à la fin d'un tour mais remet le ou les compteurs concerné(s) à 0. Le joueur peut l'utiliser de manière plus ou moins pragmatique pour éviter de gaspiller des ressources. **Canne à pêche** (rare, carte permanente) : le joueur qui la possède récolte plus de poisson à chaque tour où il y va. **Gourde** (rare, carte permanente) : le joueur qui la possède récolte plus d'eau à chaque tour où il y va. **Planche** (rare) : apporte 1 radeau entier. Noix de coco (rare), bouteille d'eau (courante), sardines (rare), sandwich (courante) : ressources d'eau et de nourriture.

Cartes pragmatiques	**Kit BBQ cannibale** (rare) : permet de transformer chaque joueur mort pendant le tour en 2 ressources de nourriture. Nous observerons l'optimisation de son utilisation. **Hache** (rare, carte permanente) : le joueur qui la possède récolte plus de bois à chaque tour où il y va. **Café** (rare) : permet de faire 2 actions en 1 tour (1 fois). Conserver la carte fait perdre 1 tour mais elle pourra être utilisée lors d'un tour qui rapporte plus ou par un joueur qui a un avantage (gourde, canne à pêche, hache). **Moulin à légumes** (rare) : transforme 2 ressources de nourriture en 2 ressources d'eau. **Nouilles chinoises** (rare) : transforme 2 ressources d'eau en 2 ressources de nourriture.
Cartes collaboratives	**Poupée vaudou** (rare) : ressuscite 1 mort. **Anti-venin** (rare) : guérit 1 malade (mordu par le serpent en allant chercher du bois – un malade ne peut pas faire d'actions pendant 1 tour, ni voter, ni utiliser ses cartes). Peut avoir un usage individualiste (garder pour soi), pragmatique (donner à un joueur / à un moment plus utile) ou collaboratif (donner au 1er qui tombe malade).
Cartes inutiles	Eau croupie, poisson pourri, vieux slip, brosse à WC, magazine minceur (courantes)…

- Questionnaires en ligne (Google Forms) : questionnaires GDMS (style de prise de décision), Big Five (personnalité) et MCQ-30 (métacognitif).
- 1 caméra (Handy Video Recorder Q2n-4K ZOOM).

22 joueurs ont été mobilisés pour la phase C de l'expérimentation, avec un seul participant par passation. Les prérequis pour les participants étaient d'être majeur et de comprendre et parler le français couramment. Pour cette phase, nous acceptions les participants qui ne connaissaient pas le jeu car celui-ci est facile à apprendre. Si le participant ne connaît pas le jeu, nous prévoyons une partie d'entraînement supplémentaire.

La mise en œuvre du protocole expérimental a nécessité deux expérimentateurs (un *meneur* et un *agent*) communiquant via un partage d'écran par visioconférence et une conversation écrite. L'*agent* joue les deux participants simulés sur Panadarchipel, *Léa* et *Thomas*, en suivant les recommandations du *meneur*, qui lui indique notamment quand démarrer et quels messages d'influence envoyer. Il se trouve dans une salle différente afin de ne pas donner d'indices qui pourraient permettre au participant de s'apercevoir qu'il simule les autres joueurs (synchronisation des frappes au clavier, etc.). Le *meneur* se trouve dans la même salle que le participant pour le diriger dans les tâches à effectuer. Il réalise également l'observation des *« cues »* manuelles dans Postdare (en visionnant l'écran de l'*agent*) et conduit l'entretien.

L'objectif de cette dernière phase -C- de l'expérimentation était de tenter de répondre aux questions de recherche (voir introduction). Pour ce faire, elle s'appuyait sur l'arbre de détection d'informations critiques obtenu lors des deux premières phases -A et B- et implémenté dans Postdare afin de déterminer en temps-réel la stratégie mise en place par le joueur à partir des signaux faibles (cues) détectés par le logiciel. De plus, nous cherchions à influencer le joueur vers la stratégie opposée à la sienne grâce aux variables identifiées au cours de l'expérimentation précédente (voir *Chapitre 3*).

Pour ce faire, l'outil numérique Panadarchipel (Figure 41) permettait d'isoler le participant tout en lui donnant l'illusion de joueur avec d'autres personnes (en réalité simulées par un expérimentateur – l'*agent*). En parallèle, l'outil Postdare implémentait l'arbre de détection d'informations critiques élaboré lors des phases A et B de l'expérimentation, ce qui permettait une sélection automatique de la plupart

des *cues* grâce à une communication entre les deux logiciels (Figure 44). Certaines cues n'ont pas pu être automatisées et nécessitaient une sélection manuelle par un expérimentateur (le *meneur*), notamment celles impliquant une interprétation de l'état du plateau de jeu et des probabilités de survie (cf. Figure 42).

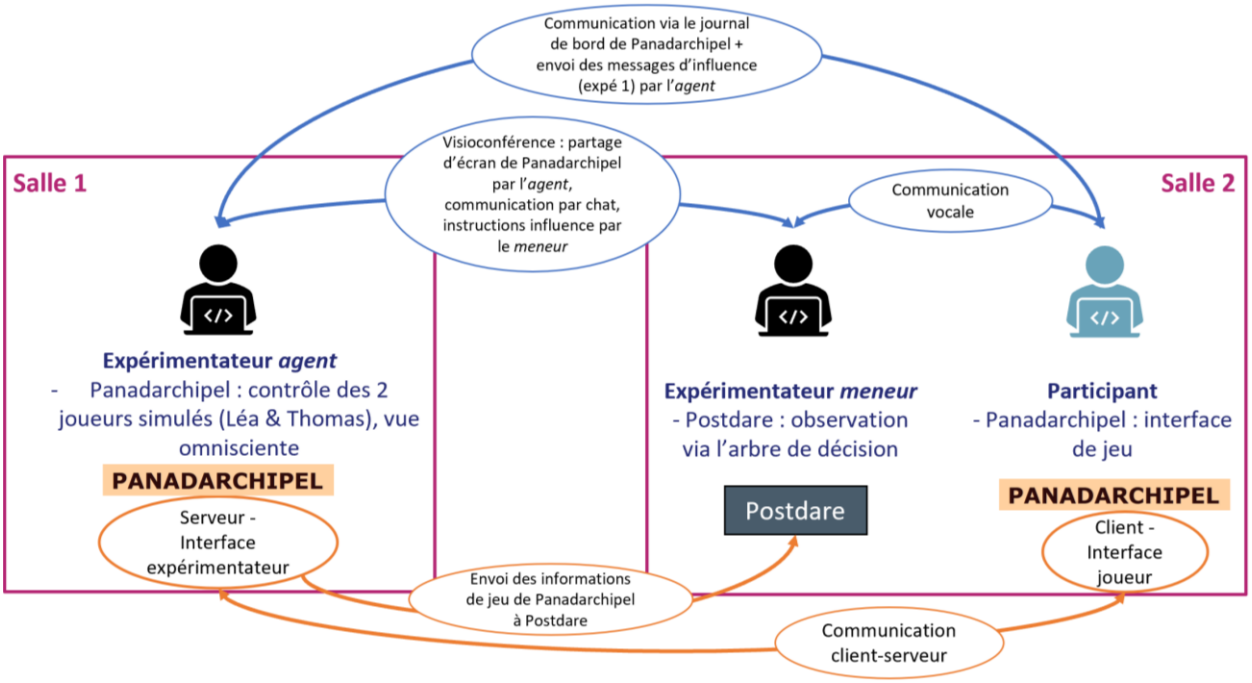

Figure 44 : Protocoles de communication entre les parties prenantes de la phase C et les logiciels utilisés

Les 12 étapes du protocole de la phase C sont présentées dans le Tableau 20.

Tableau 20 : Protocole détaillé de la phase C en 12 étapes

Étape	Données à récolter	Durée
0. Consignes, description du déroulement et signature du consentement	/	3 min
1. Questionnaires Big Five, MCQ-30, GDMS	Traits de personnalité, structures des croyances métacognitives, profil décisionnel du joueur	10-15 min
2. Rappel des règles du jeu et entraînement – prise en main de l'outil Panadarchipel (Galèrapagos numérique) avec une partie d'entraînement	/	10-20 min
3. Questionnaire main initiale idéale	4 cartes de la main initiale idéale du joueur et explications	2 min
4. Partie 1 : observation de la stratégie via Postdare	Décisions prises par le joueur (Postdare), verbatims	20-30 min
5. Questionnaire main initiale idéale, questionnaire de ressenti sur la partie	4 cartes de la main initiale idéale du joueur et explications. Dynamique de jeu ressentie par le joueur et ses objectifs, moments décisifs pendant la partie.	5 min

Étape	Données à récolter	Durée
6. Partie 2 : observation de la stratégie via Postdare + tentatives d'influence	Décisions prises par le joueur (Postdare), verbatims	20-30 min
7. Questionnaire de ressenti sur la partie	Dynamique de jeu ressentie par le joueur et ses objectifs, moments décisifs pendant la partie.	3 min
8. Partie 3 : observation de la stratégie via Postdare + tentatives d'influence	Décisions prises à chaque étape (saisie dans Postdare), communication écrite du joueur (journal de bord de la partie) et éventuellement orale.	20-30 min
9. Questionnaire main initiale idéale + questionnaire de ressenti sur la partie	4 cartes de la main initiale idéale + explications. Dynamique de jeu ressentie par le joueur et ses objectifs, moments décisifs pendant la partie.	5 min
10. Questionnaire démographique	Âge, niveau d'études, genre, nationalité, niveau d'expertise au Galèrapagos, habitude de jouer à des jeux de société.	20 min
11. Questionnaire MIST	Sensibilité de chaque joueur à la mésinformation (score total), scores de faux positifs et faux négatifs.	5 min
12. Entretien d'évaluation de la stratégie (en se basant sur le journal de bord de la partie 3) et de l'influence	Vision du participant sur sa propre stratégie, explication des actions décisives prises dans le jeu. Déterminer si et comment le joueur a perçu les tentatives d'influence.	25-30 min
Total	**La durée totale dépend de la durée des parties (généralement 20 à 30min).**	**2h30-3h**

Étape 0 : Nous commençons par donner aux participants les consignes de l'expérimentation, en décrire le déroulement, puis leur faire signer un formulaire faisant état de leur droit de retrait, des mentions RGPD et recueillant leur consentement explicite.

Étape 1 : Nous demandons au participant de compléter les mêmes questionnaires qui ont été utilisés lors de la phase B, afin de récolter plus d'informations sur le profil du joueur et éventuellement d'utiliser ces informations pour expliquer ses choix stratégiques et confirmer ou infirmer le modèle de prédiction de la stratégie ébauché lors de la *Phase B* (modèle présenté dans le paragraphe *4.3.2.3 Modèle de prédiction des stratégies à partir des réponses aux questionnaires*). Les questionnaires de personnalité, métacognition et style de prise de décision sont remplis en ordre contrebalancé pour éviter les effets d'entraînement ou de fatigue. Ils sont complétés avant les parties de Galèrapagos afin de s'assurer que la tâche ne subisse aucun effet du déroulement des parties sur les réponses à ces questionnaires.

Étape 2 : L'expérimentateur *meneur* lit les règles du jeu au participant, y compris s'il est expérimenté afin de s'assurer que chacun démarre le jeu avec les mêmes règles. Ainsi, sont rappelés : le déroulement d'un tour de jeu, les différentes actions possibles et leur déroulement (conserver une carte piochée, récolter du bois, récolter de la nourriture ou récolter de l'eau), les types de cartes disponibles, le déroulement des votes ou éliminations par le revolver, les effets d'une morsure de serpent (qui est un risque pris par le joueur lorsqu'il récolte du bois), le nombre de tours de jeu avant l'arrivée possible de la tempête, les conditions pour quitter l'île. L'adaptation des règles spécifiques à l'expérimentation (pioche systématique de carte) est énoncée et l'interface est présentée, notamment avec la présence de cartes rares ou

fréquentes (qui ont des chances d'être piochées 2 fois au cours de la même partie ou non). La seule indication donnée au joueur pour gagner la partie est de *« quitter l'île »*, ce qu'il peut interpréter comme partir seul ou partir avec les autres joueurs.

Le participant joue ensuite une première partie de Galèrapagos sur l'outil Panadarchipel (partie raccourcie si le joueur est déjà expérimenté : la tempête arrive alors au 5e tour au lieu d'arriver entre le 7e et le 9e tour, sinon d'une durée normale). Si le joueur n'est pas familier avec le jeu et n'a pas totalement intégré les règles, il fera ensuite une seconde partie d'entraînement raccourcie. Au cours de la ou des partie(s) d'entraînement, l'expérimentateur *agent* simulant les deux autres participants (joueurs imaginaires *Léa* et *Thomas*) a pour consigne d'adopter une posture de jeu neutre.

NB : La partie d'entraînement comme les parties jouées par la suite sont préparées à l'avance (tirages de cartes et météos – voir *Annexe 2*). La partie d'entraînement est la même pour tous les joueurs (2 parties d'entraînement possibles si le joueur est débutant). Il y a ensuite 3 parties préparées différentes pour les étapes 4, 6 et 8, qui sont présentées dans un ordre contrebalancé pour chaque participant.

Étapes 3, 5, 7 et 9 : Nous demandons au participant de remplir trois fois le questionnaire indiquant sa main initiale idéale au jeu de Galèrapagos : il doit sélectionner parmi une liste comportant les photos de toutes les cartes du jeu, les 4 cartes qu'il aimerait avoir en main en commençant une partie. Ce questionnaire nous donne des indices sur la stratégie du participant, les cartes ayant une connotation collaborative, pragmatique, individualiste ou irrationnelle (cf. Tableau 19). Nous donnons à chaque main initiale idéale un caractère qui dépend des cartes sélectionnées, avec la possibilité d'établir des états intermédiaires (par exemple, « collaboratif / pragmatique ») si la main initiale sélectionnée permet plusieurs styles stratégiques. Un champ texte final dans le questionnaire demande au joueur d'expliquer ses choix de cartes, en espérant qu'il explicite sa stratégie. L'explication sert à affiner l'interprétation de la main initiale idéale, et donc de la tendance stratégique du joueur. Ce questionnaire est complété pour la première fois avant la partie d'entraînement pour les joueurs expérimentés, ou juste après l'entraînement pour les joueurs qui ne connaissaient pas le jeu. L'objectif est de collecter un indice supplémentaire sur la stratégie du participant avant qu'il risque d'être influencé par les autres « participants » et leur manière de jouer. La deuxième fois, il est complété après la première partie observée, donc avant toute tentative d'influence. La troisième fois intervient après la dernière partie jouée, donc après les tentatives d'influence. Nous espérons ainsi corroborer ou infirmer d'autres observations (Postdare, questionnaires de ressenti sur la partie et entretien).

Quant au questionnaire de ressenti sur la partie, il est rempli après chaque partie observée, avec pour objectif de récolter le ressenti du joueur sur les actions des autres joueurs et ses propres choix et objectifs au cours du jeu. Il est composé des questions suivantes :

- *Avez-vous ressenti une dynamique de groupe particulière dans cette partie ? Collaborative / Individualiste / Ambigüe / Autre (à compléter)*
- *Avez-vous senti à certains moments que la partie prenait une tournure décisive ? Si oui, pourquoi ? (réponse libre)*
- *Aviez-vous des objectifs en débutant cette partie ? Si oui, lesquels ? (réponse libre)*
- *Vos objectifs ont-ils changé au cours de la partie ? Si oui, pourquoi ? (réponse libre)*
- *Votre manière de jouer a-t-elle changé au cours de la partie ? Si oui, pourquoi ? (réponse libre)*

En cas de doute sur la stratégie initiale du joueur via Postdare (joueur avec un équilibre entre les actions individualistes et collaboratives ou très majoritairement pragmatique), les premiers résultats de ces deux questionnaires sont utilisés pour déterminer vers quelle stratégie l'influencer.

Étape 4 : Les trois parties jouées (étapes 4, 6 et 8) suivent un ordre prédéterminé en ce qui concerne les météos de chaque tour et la distribution des cartes (voir *Annexe 2*).

Le joueur a pour consigne de *« quitter l'île »*, sans préciser s'il doit gagner seul ou en groupe, afin de lui permettre d'interpréter les règles suivant ses préférences personnelles. Chacune de ces parties a été conçue pour permettre au participant d'exprimer différents types de stratégies. Aux tours 3, 5 et 7, le joueur aura notamment un choix décisif à faire au niveau d'une carte à conserver : soit une carte foncièrement individualiste, soit une carte foncièrement collaborative, qui donnera alors un indice sur sa stratégie.

Lors de la première partie jouée, nous observons uniquement la stratégie via le logiciel Postdare qui implémente l'arbre de détection d'informations critiques. Si la plupart des cues sont sélectionnées automatiquement, certaines ne pourront pas être automatisées pour des raisons de complexité ou parce qu'elles nécessitent une part d'interprétation de la situation de jeu (chances de partir à plusieurs, conversations observées entre les joueurs etc.). Ces cues sont tout de même décrites précisément et le plus objectivement possible.

Au cours de cette partie et des suivantes, l'expérimentateur *agent* a pour consigne d'adopter une posture collaborative pour *Léa* et plus individualiste pour *Thomas*, avec des consignes précises sur les actions à effectuer au cours de chaque partie (quelles cartes garder ou défausser, quand utiliser quelle carte…). L'objectif est de donner au participant deux référentiels opposés afin de limiter le risque qu'il se range au comportement du groupe par mimétisme et qu'il puisse choisir sa propre stratégie plus librement.

Pour les trois parties observées (étapes 4, 6 et 8), nous avons préparé différents messages automatiques à envoyer au joueur en ordre aléatoire : cinq messages neutres et quatre messages d'influence, ayant chacun deux versions, l'une orientant vers une stratégie individualiste et l'autre orientant vers une stratégie collaborative (Tableau 21). Ces messages seront envoyés au participant via le chat de la partie, sous un format différent des messages des autres joueurs et sous le nom de *« protips »* ; le participant est prévenu avant la première partie observée qu'il recevra des messages *« provenant d'un filtre qui cherche sur les réseaux sociaux et des forums des conseils qui semblent pertinents pour s'améliorer au jeu de Galèrapagos, et qui enverra des messages aléatoirement au cours des parties, dans l'objectif de tester ce filtre »*. Ces messages sont en réalité envoyés par l'expérimentateur *agent* qui joue les deux participants fictifs. Nous avons fait le choix de présenter les messages d'influence comme un facteur externe *« aléatoire »* afin qu'ils n'apparaissent pas au joueur comme étant dépendants de la tâche ou de ses choix de jeu.

Au cours de la première partie observée (étape 4), seuls deux messages dits « neutres » sont envoyés, par ordre contrebalancé pour chaque participant : ces messages ont été rédigés manuellement pour apparaître comme venant des réseaux sociaux, étant soit des astuces (1 et 2), soit des publicités (3 et 4), soit des commentaires (5) (voir Tableau 21). Les messages dits « d'influence » (utilisés seulement lors des étapes 6 et 8) ont été conçus comme ceux de l'expérimentation 1 : une première version a été rédigée manuellement avant d'utiliser l'IA générative Chat-GPT pour y introduire les facteurs de crédibilité souhaités.

Tableau 21 : Messages neutres et d'influence envoyés au cours des parties observées et présentés comme provenant d'un « filtre ». Entre parenthèses en italique sont indiqués les facteurs issus de l'expérimentation 1 qui ont été utilisés dans le message concerné

<table>
<tr><td>Messages neutres</td><td colspan="2">1. Le saviez-vous ? Le Kit BBQ Cannibale vous permet de gagner 2 rations de nourriture pour chaque joueur mort au cours du tour. En plus des rations économisées à la fin du tour.. Ensuite, vous pourrez toujours ressusciter quelqu'un avec la Poupée Vaudou !

2. Mon astuce : si météo 2 pendant le tour et j'ai les nouilles chinoises, il vaut mieux chercher l'eau et faire de la nourriture avec vu qu'on peut tomber sur la boule 1 poisson

3. Galerapagos lovers, vous connaissez les naufragés du titanic ? Vous avez des retours à faire dessus ?

4. Les variantes : à retrouver sur le site de Galèrapagos, d'autres versions des règles pour pimenter un peu vos parties et varier les plaisirs !

5. On peu pas commencer la partie avec + de ressource ? C trop dur</td></tr>
<tr><td rowspan="2">Messages d'influence</td><td>Orienter vers une stratégie collaborative :</td><td>Orienter vers une stratégie individualiste :</td></tr>
<tr><td>6. @Galerapagos-officiel : notre dernière étude menée par nos spécialistes en théorie des jeux montre que les chances de survie, même individuelles, sont plus élevées quand on joue de manière collaborative en se mettant d'accord sur les actions à prendre. (sources)

7. On réussit mieux en prenant les décisions collectivement. Si on décide ensemble de ce qu'on peut faire pendant le tour et on parle de nos ressources, on peut mieux s'organiser pour optimiser les actions, avancer et survivre. Par exemple, si je pioche la gourde alors que la météo est 2 ou 3, il est plus judicieux de la donner à quelqu'un qui joue après moi pour qu'il s'en serve dès ce tour (exemples et détails)

8. Les chances de survie, même individuelles, sont plus élevées quand on joue de manière collaborative. On a alors plus de chances maximiser les ressources, survivre ensemble et gagner la partie en quittant tous l'île avant la tempête (répétitions)</td><td>10. @Galerapagos-officiel : notre dernière étude menée par nos spécialistes en théorie des jeux montre que les chances de survie sont plus élevées lorsqu'on élimine les autres en fin de partie et garde des ressources dans son jeu. (sources)

11. Parfois en jouant un peu solo on peut partir seul ou à 2 très tôt ! Surtout que si quelqu'un meurt on récupère ses ressources, ou même on le transforme en nourriture avec le kit cannibale. On a alors assez à manger pour finir le bois vite fait. D'ailleurs, perso je garde au moins des ressources dans mon jeu si j'ai pas la conque, au cas où on vote pour m'éliminer (exemples et détails)

12. On a plus de chances de survivre à la fin si on pense un peu à soi en gardant des ressources et des cartes défensives au cas où. On a alors plus de chances de maximiser ses propres ressources, survivre et de gagner la partie en quittant l'île avant la tempête (répétitions)</td></tr>
</table>

	9. L'approche collaborative renforce considérablement les perspectives de survie, même pour chaque individu. Cette méthode permet une gestion plus efficace des ressources, améliore les chances de survie collective et augmente la probabilité de quitter l'île avant l'arrivée de la tempête. *(SMOG)*	13. Les perspectives de survie en fin de partie s'améliorent si l'on préserve certaines ressources et cartes défensives pour soi-même. Cette stratégie permet d'optimiser l'utilisation de ses propres ressources, d'augmenter ses chances de survie et de remporter la victoire en quittant l'île avant l'arrivée de la tempête. *(SMOG)*

Étapes 6 et 8 : Pendant les deux dernières parties observées, l'expérimentateur *agent* suit les mêmes consignes qu'au cours de l'étape 4 (première partie observée). Seuls les messages du « filtre » changent : au cours de l'étape 6 (2e partie, observée et influencée) sont envoyés 1 message neutre, 1 message d'influence puis 1 message neutre, et au cours de l'étape 8 (3e partie, observée et influencée), sont envoyés 1 message d'influence, 1 message neutre puis 2 messages d'influence, toujours avec un ordre contrebalancé sur l'ensemble des messages envoyés pendant les 3 parties. Les messages dits « d'influence » sont sélectionnés en fonction de la stratégie du joueur observée via Postdare : si la stratégie dominante du participant est individualiste, il recevra des messages l'invitant à devenir plus collaboratif, et inversement. Il peut arriver qu'aucune stratégie dominante ne se dégage via Postdare, ou que la stratégie dominante soit pragmatique avec une part équilibrée d'individualisme et de collaboration. Dans ce cas, les expérimentateurs s'appuient sur les réponses aux questionnaires de ressenti de la partie et de la main initiale idéale pour déterminer vers quelle stratégie orienter le joueur : si ceux-ci révèlent une tendance stratégique générale individualiste, le participant sera influencé vers une stratégie collaborative, et inversement. L'irrationalité est restée jusqu'ici marginale, mais si un joueur est majoritairement irrationnel d'après Postdare, nous ferons le choix de l'écarter lors de l'analyse des résultats car ce type de participant est imprévisible, souvent par méconnaissance du jeu ou par désintérêt.

Nous avons fait le choix de tenter d'influencer le participant seulement au cours de la tâche plutôt que d'avoir la moitié des participant qui commencent avec la modalité « influence » et terminent sans et l'autre moitié qui fait l'inverse. Ce choix s'explique par deux raisons. Premièrement, nous savons quand l'influence commence mais pas lorsqu'elle s'arrête, si elle fonctionne elle pourrait continuer à avoir un effet durable jusqu'à la fin de l'expérimentation même pour les participants qui commencent « avec influence ». Deuxièmement, en démarrant l'expérimentation, nous ne savons pas quelle est la stratégie du joueur avant d'avoir observé via Postdare une première partie sans influence.

Les messages d'influence sont élaborés sur la base des résultats issus de l'expérimentation 1 (voir *Chapitre 3*). En effet, cette expérimentation a montré que la présence dans un message textuel de sources, d'exemples ou détails, de répétitions et de longues phrases avec des mots de trois syllabes ou plus (score SMOG élevé) confèrent à ce message plus de crédibilité. Nous avons donc conçu des messages destinés à orienter les joueurs vers un comportement plus individualiste (« vers individualiste ») ou plus collaboratif (« vers collaboratif ») sur la base de chacun de ces facteurs. L'objectif est de renforcer la capacité d'influence des messages. Lors de l'entretien, nous chercherons à identifier les facteurs qui sont particulièrement efficaces dans ce contexte.

Étapes 10 et 11 : Les questionnaires démographique (étape 10) et MIST (étape 11) sont remplis par le participant après les parties, afin que l'expérimentateur *meneur* qui conduit l'entretien puisse consulter les réponses aux différents questionnaires de main initiale idéale et de ressenti, et se préparer à l'échange. Le questionnaire démographique a pour objectif de récolter des informations permettant une approche

de statistiques descriptives du panel de participants, ainsi que de recueillir leur niveau d'expertise au Galèrapagos et dans les jeux de société en général. Ces deux derniers items constituent des indices pouvant expliquer le déroulement des parties et certains choix d'actions ou de stratégies. Le questionnaire MIST vise à recueillir le niveau de sensibilité à la désinformation, afin de vérifier s'il y a un lien entre les résultats de ce questionnaire et la tendance à se laisser influencer par les messages du « filtre ».

Étape 12 : L'expérimentation se termine avec un entretien enregistré en vidéo, visant à recueillir la vision du joueur sur les parties et sa propre stratégie. Les points abordés sont les suivants :

- Les choix majeurs effectués au cours des deux premières parties et qui ont été notés par l'expérimentateur *meneur* (par exemple l'utilisation d'une carte individualiste « agressive » envers les autres joueurs, un sacrifice au profit des autres joueurs, etc.).
- L'expérimentateur *meneur* et le participant reviennent ensemble sur la dernière partie jouée en visualisant le journal de bord qui liste toutes les actions et conversations, ainsi que l'état du plateau au début de chaque tour de jeu. Le participant explique alors :
    - Ses choix d'actions : pour vérifier l'interprétation de ces actions en tant que *cues* dans Postdare.
    - Ses interactions significatives avec les autres joueurs : afin de mieux comprendre les interactions, les influences potentielles et les intentions du joueur.
    - Son avis sur les messages « protips », s'il les a lus, pris en compte, et ce qu'il en a pensé : dans le but d'évaluer leur perception et la présence ou absence d'influence.
    - Ses objectifs ou pensées au cours du jeu : pour vérifier l'analyse de la stratégie par Postdare.
- L'expérimentateur demande ensuite au participant s'il pense que les actions ou paroles des autres joueurs ont pu avoir un impact sur sa manière de jouer, afin de vérifier si les choix de l'expérimentateur qui simule Léa et Thomas ont pu introduire un biais.
- L'expérimentateur revient sur les objectifs du joueur au cours de la partie, si ce point n'a pas déjà été abordé lors des questions précédentes, afin d'évaluer sa vision sur sa propre stratégie.
- L'expérimentateur montre au participant l'ensemble des messages « protips » qu'il a reçus au cours des trois parties et lui demande s'il en a trouvé certains intéressants, et si oui, lesquels étaient les plus intéressants. L'objectif est alors de mesurer à la fois l'efficacité des protips en fonction des facteurs de l'expérimentation 1 associés (présence de sources, d'exemples et détails, d'un langage peu clair ou de répétitions), et si le participant a été influencé par ces textes. Lors de l'analyse de l'entretien, la réponse donnera lieu à une mesure de l'intérêt porté à chaque « protip » entre 0 (« pas d'intérêt ») et 3 (« très intéressant »).
- Enfin, l'expérimentateur demande au participant son avis sur l'outil de Galèrapagos numérique, afin de déterminer si certaines choses ont pu le gêner dans son jeu et dans l'application de sa stratégie de prédilection et pour pouvoir améliorer l'outil par la suite en cas de réutilisation.

4.3 Résultats

4.3.1 Résultats de la phase A

Comme décrit dans le paragraphe *4.2.2.1 Phase A : Croisement de regards d'experts*, le focus group a rassemblé à plusieurs reprises 5 experts dont 2 femmes et 3 non experts du jeu de Galèrapagos, tous diplômés de Bac+5 à Bac+8. Ils ont été confrontés à diverses activités autour du jeu.

Le groupe d'experts a distingué quatre styles stratégiques possibles, qui correspondent aux quatre CCIR de l'arbre de détection d'informations critiques (Figure 50 – p. 142) : Individualiste (privilégie sa propre survie), Pragmatique (fait survivre le groupe si possible, sinon lui-même), Collaboratif (privilégie la survie du groupe complet), Irrationnel (stratégie illogique avec des actions qui ne favorisent ni sa propre survie, ni celle du groupe). Une soixantaine de cues ont été identifiées, séparées en 24 triggers. Pour chaque CCIR, soit chaque style stratégique, il existe des triggers équivalents qui correspondent aux catégories suivantes (voir un extrait Figure 50) :

- Partage des informations avec les autres joueurs : dévoilement de ses propres ressources, mensonge ou dissimulation à propos de ses cartes, etc.
- Actions du tour : choix de récolter de l'eau, de la nourriture ou du bois, ou de conserver une carte.
- Décisions de survie : votes, don de ressources présentes dans son jeu, etc.
- Utilisation des cartes : utilisation d'une carte ressource, défensive, agressive, ou pour aider un ou plusieurs autres joueurs.
- Débats : communication avec les autres joueurs, incite les autres joueurs à agir dans un sens ou un autre, demande certaines cartes, etc.
- Comportement général : mode de communication, langage utilisé, émotions, etc.

Un poids est associé à chaque cue. Une cue peut être rattachée à plusieurs triggers liés à des CCIR différents, avec un poids différent (voir Figure 50). Cela permet, par exemple, de modéliser le fait que l'utilisation du *taser* a une probabilité importante d'être une action individualiste, mais peut aussi être pragmatique avec une probabilité plus faible cependant.

4.3.2 Résultats de la phase B

4.3.2.1 Caractéristiques des participants

La phase B de cette expérimentation s'est déroulée pendant une journée entière, avec 3 passations rassemblant chacune 4 participants. Cette phase s'est donc déroulée avec un total de 12 participants, dont 4 femmes et 8 hommes âgés de 22 à 24 ans. Il s'agissait d'élèves français en école d'ingénieurs.

4.3.2.2 Validation et correction de l'arbre de détection d'informations critiques pour l'évaluation des stratégies

La phase B nous a permis de valider l'existence de 4 styles stratégiques (individualiste, pragmatique, collaboratif, irrationnel), tel que nous l'avions anticipé lors de la phase A (Figure 45). Cependant, le style

irrationnel est peu présent et si certaines actions irrationnelles sont observées, elles restent marginales et ne constituent jamais un style dominant.

La classification du style stratégique dans la main initiale idéale est basée sur le caractère stratégique des cartes sélectionnées (individualiste, pragmatique, ressource collective…) et sur l'interprétation du texte d'explication des choix rédigé par le joueur. La classification du style stratégique au cours de la partie repose sur les actions choises au cours du jeu, les cartes conservées ou défaussées et les interactions et partages d'informations avec les autres joueurs. Chez certains joueurs, le style stratégique détecté au cours de la partie peut différer de celui de la main initiale idéale en fonction du contexte de la partie : opportunités, cartes en main, état du plateau de jeu, ajustement au comportement des autres joueurs…

Ainsi, parmi les 12 participants, seuls 2 étaient individualistes (Figure 45 – couleur grise), ce qui se traduisait à la fois par une main initiale idéale agressive et un style de jeu individualiste au cours de la partie observée. Parmi les 10 autres, 4 étaient collaboratifs (Figure 45 – couleur verte), 3 pragmatiques et 3 entre pragmatique et collaboratif (Figure 45 – couleur bleue) au cours de la partie. Leurs mains initiales idéales étaient également collaboratives ou pragmatiques mais ne correspondaient pas toujours au comportement observé au cours du jeu : la délimitation entre ces deux styles stratégiques semble floue ou facile à franchir par les joueurs.

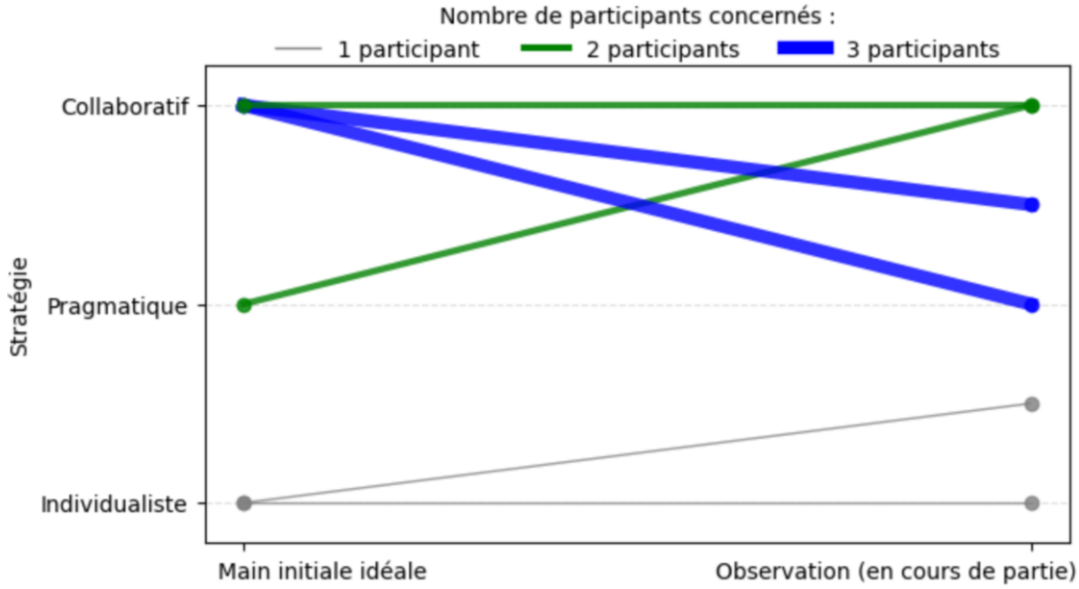

Figure 45 : Comparaison des tendances stratégiques entre le questionnaire de main initiale idéale et l'observation au cours de la partie (12 participants)

Certaines hypothèses formulées lors de la phase A en ce qui concerne les justifications des actions des joueurs et leurs liens avec les stratégies ont été revues, nous conduisant à reformuler ou à supprimer des cues. D'autres ont été validées et nous avons fait de nouvelles observations qui ont mené à l'ajout de cues supplémentaires dans l'arbre de détection d'informations critiques. Nous les présentons ci-après.

Conserver ou défausser une carte :

Hypothèse 1 : *Conserver une carte agressive (revolver, cartouches, somnifères…) traduit une stratégie individualiste*. Nous pouvons constater sur la Figure 46 que les joueurs individualistes ont effectivement conservé toutes les cartes agressives ou défensives. Au contraire, seul un joueur pragmatique a conservé une fois une carte agressive, la plupart des joueurs collaboratifs ou pragmatiques défaussant les cartes agressives et défensives. L'hypothèse 1 est donc vérifiée sur l'échantillon évalué.

Hypothèse 2 : *Conserver une carte utile pour la communauté (planche, canne à pêche, hache, gourde) traduit une stratégie collaborative*. La Figure 46 montre que tous les joueurs conservent les cartes en question, sauf une erreur d'un joueur pragmatique. L'hypothèse 2 n'est donc pas vérifiée, cette action ne permet pas de discriminer les stratégies.

Hypothèse 3 : *Tous les joueurs défaussent les cartes valant une seule ressource (bouteille d'eau, sandwich, eau croupie, poisson pourri) et préfèrent faire une action qui peut leur rapporter plus de ressources*. Nous pouvons voir sur la Figure 46 qu'aucun joueur n'a conservé une des cartes en question. L'hypothèse 3 est donc vérifiée : ces cartes peuvent être apparentées à des cartes inutiles, l'action de les conserver ne permet pas de discriminer les 3 styles stratégiques observés lors de la phase B. Elle pourrait éventuellement traduire une stratégie irrationnelle. Nous observons que c'est également le cas des allumettes (qui correspondent à une seule ressource lorsqu'elles sont associées au poisson pourri ou à l'eau croupie) et de l'anti-venin (qui revient presque à échanger une action contre une autre).

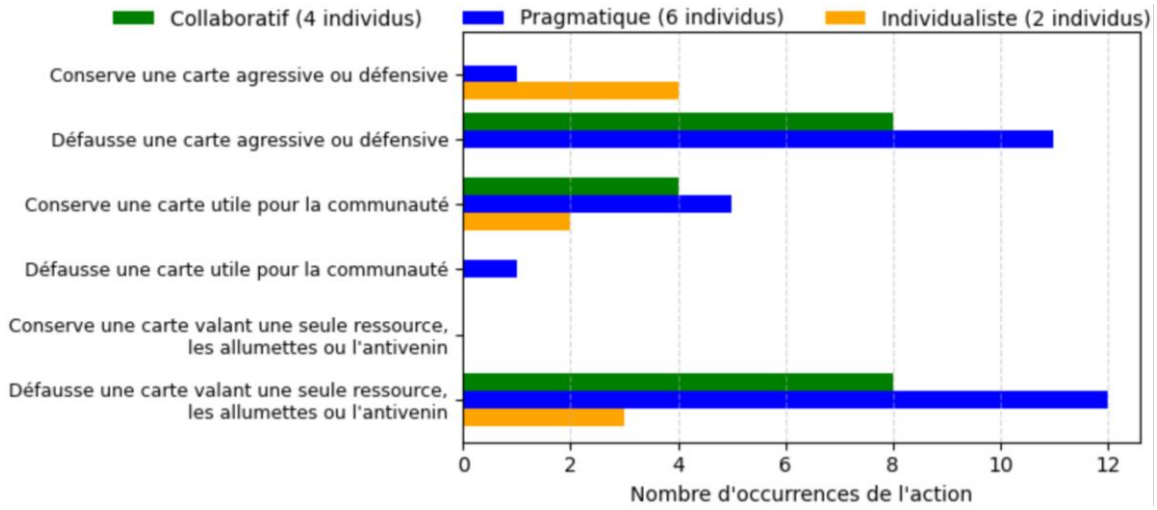

Figure 46 : Choix de conserver ou défausser une carte selon le style stratégique dominant

Dévoiler une carte :

Hypothèse 4 : *Dévoiler une carte agressive alors que ce n'est pas pertinent pour le collectif (pas besoin par exemple d'éliminer un joueur rapidement pour garantir la survie du plus grand nombre) traduit un comportement individualiste ou agressif*. Nous pouvons observer sur la Figure 47 que cette action dépend avant tout des circonstances : en effet, ici elle a plutôt été observée lorsqu'il était nécessaire d'éliminer un joueur pour la survie du groupe. Elle a aussi été observée chez des joueurs collaboratifs et pragmatiques, presque tous membres du premier groupe de 4 joueurs, qui ont chacun dévoilé leurs cartes agressives à la fin du jeu par mimétisme. L'hypothèse 4 n'est donc pas vérifiée ici : la plupart des joueurs ne dévoilent une carte agressive que si leur propre survie ou celle du plus grand nombre en dépend, ou si l'ambiance et la communication dans le groupe les y invite. En revanche, nous avons observé chez les joueurs individualistes des comportements verbaux provocateurs qui pouvaient laisser planer un doute sur leur possession de cartes agressives sans les dévoiler ouvertement. Ces comportements ne peuvent pas être inscrits comme cues automatiques dans l'arbre de détection d'informations critiques de Postdare mais pourront servir d'indices à détecter et saisir manuellement.

Hypothèse 5 : *Cacher ses cartes défensives traduit une stratégie individualiste*. Cette hypothèse n'a pas été vérifiée : en effet, la Figure 47 montre que tous les joueurs ont caché leurs cartes défensives du début à la fin de la partie, indépendamment de leur type de stratégie.

Hypothèse 6 : *Les joueurs individualistes ne dévoilent leurs ressources que lorsque c'est nécessaire et les joueurs collaboratifs les dévoilent spontanément*. Nous pouvons observer sur la Figure 47 que les joueurs individualistes n'ont effectivement dévoilé leurs ressources que lorsque c'était nécessaire pour la survie du plus grand nombre. En revanche, certains joueurs collaboratifs ou pragmatiques le font aussi, bien que seuls des joueurs collaboratifs ont dévoilé leurs ressources spontanément. Plusieurs joueurs collaboratifs et pragmatiques les ont également dévoilées par mimétisme ou au dernier tour, lorsqu'il n'était plus justifiable de les conserver pour plus tard.

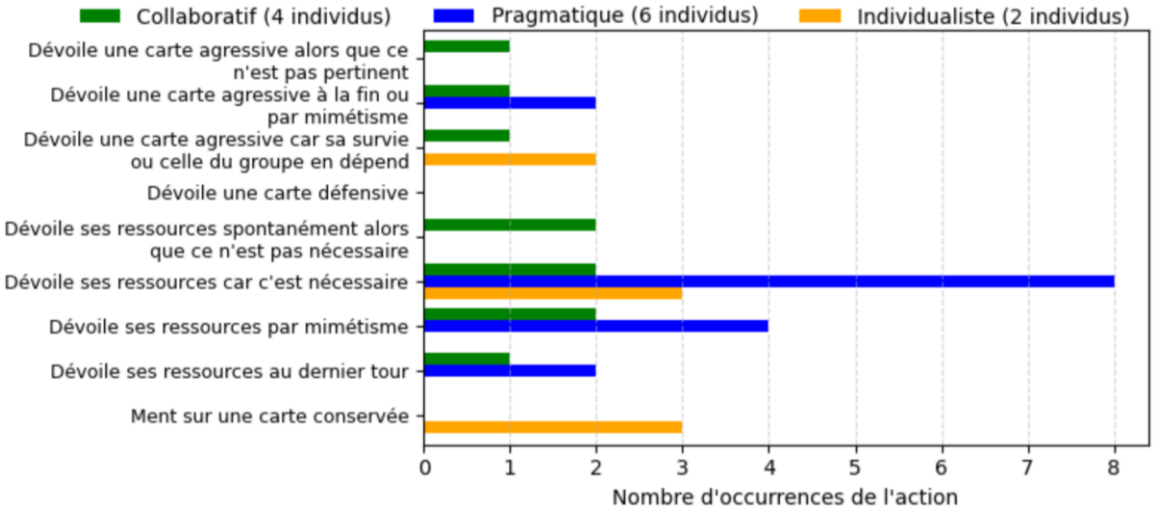

Figure 47 : Choix de dévoiler une carte selon le style stratégique dominant

Jouer une carte :

Hypothèse 7 : *Les cartes individualistes sont jouées uniquement par des joueurs individualistes*. Nous pouvons voir sur la Figure 48 que ça a effectivement été le cas pendant les parties jouées. Cependant, au cours de l'entretien, certains joueurs pragmatiques ont indiqué que s'ils avaient eu en main les cartes nécessaires pour éliminer un autre joueur, ils les auraient utilisées lorsque les circonstances ne permettaient pas de sauver le groupe complet et qu'un sacrifice permettait de faciliter la survie du plus grand nombre.

Hypothèse 8 : *Tous les joueurs jouent des cartes collaboratives et des cartes pragmatiques*. En effet, durant les parties observées, tous les joueurs ont utilisé les cartes qui servaient le groupe (Figure 48). Néanmoins, certains joueurs individualistes et pragmatiques ont exprimé lors de l'entretien la volonté de conserver ces cartes pour eux-mêmes lorsque c'était possible, alors que les joueurs collaboratifs les donnaient plus spontanément à la communauté.

Observation 1 : Nous avons observé 3 fois un don de carte permanente (gourde, canne à pêche ou hache) pour aider la communauté, par des joueurs individualistes ou collaboratifs indépendamment (action pragmatique – cf. Figure 48). Cette action a été influencée par l'insistance des autres joueurs pour l'un des participants individualistes concernés.

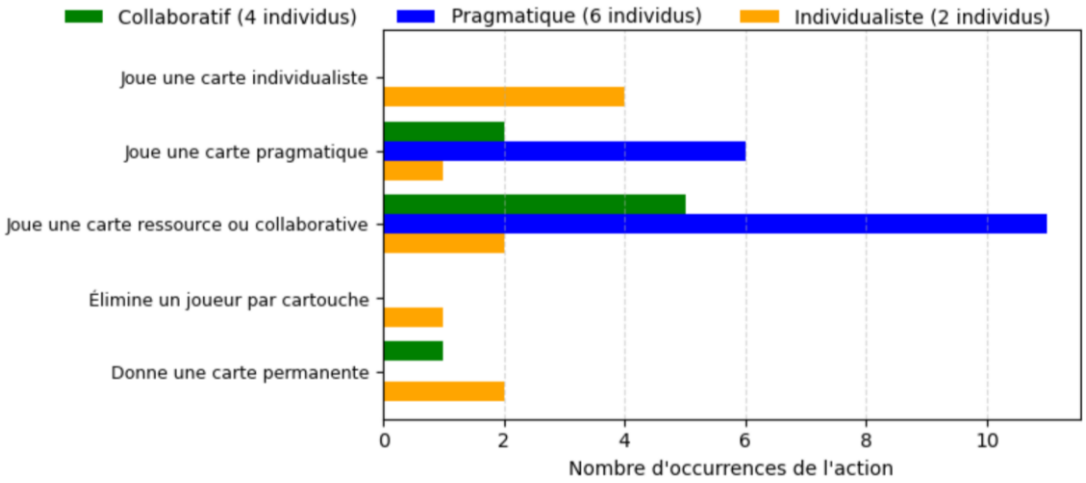

Figure 48 : Choix de jouer une carte selon le style stratégique dominant

Éliminer un joueur :

Observation 2 : L'élimination se fait par nécessité, avec l'association d'une carte *revolver* et une carte *cartouche* lorsque c'est possible, sinon par vote. Plusieurs groupes ont exprimé la volonté d'éliminer un joueur avec une cartouche afin d'économiser des ressources, mais seul le groupe qui comptait des joueurs individualistes parmi ses membres a pu mettre en place cette stratégie, tous les autres joueurs ayant défaussé les cartouches piochées.

Observation 3 : Le choix lors des votes s'est fait pour chacun d'abord par rôle dans le jeu (préserver les joueurs utiles pour la suite), puis par affinité, puis en fonction du rôle que chacun a eu avant le moment du vote (récompenser les actions les plus utiles au collectif et la collaboration), et enfin au hasard.

Observation 4 : Le seul joueur (individualiste) visé par une cartouche au cours d'une partie a essayé de se défendre (plaque de tôle).

Choisir une action (eau, nourriture ou bois) :

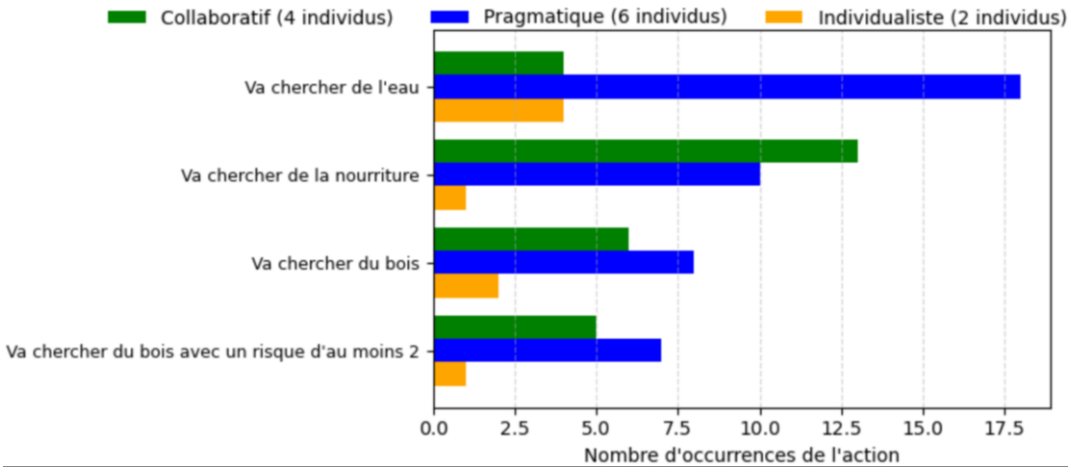

Figure 49 : Choix d'actions de jeu selon le style stratégique dominant

Hypothèse 9 : *Les joueurs individualistes évitent de prendre des risques en allant chercher du bois*. Cette hypothèse n'a pas pu être vérifiée ni écartée. Les choix d'actions semblent être avant tout expliqués par les cartes permanentes possédées apportant un avantage pour la collecte (hache, canne à pêche, gourde – voir Tableau 19) et par la situation (manque de ressources, état de la météo, avancement des radeaux,

approche de la tempête). Nous ne pouvons donc pas conclure sur la validité de cette hypothèse en l'absence des cartes permanentes et notamment de la hache.

Observation 5 : Presque tous les joueurs prennent un risque de 2 pour la collecte du bois (ce qui équivaut à un risque de 1/3 d'être mordu par le serpent – cf. Figure 49). Cela peut s'expliquer par le fait que ce niveau de risque a été énoncé comme étant le risque optimal au cours du tournoi auquel les joueurs de cette phase B ont participé au préalable. Les seules exceptions sont observées en fin de partie, lorsque les radeaux sont presque terminés et afin de réduire le risque de tomber malade.

Communiquer et partager des informations :

Observation 6 : Nous observons des tentatives de persuasion des joueurs les uns sur les autres via leur communication verbale au cours du jeu, notamment lorsqu'il s'agit de débattre des actions que chacun doit effectuer à son tour. L'ambiance de la partie a également un impact sur certains choix stratégiques des joueurs, d'après ce qu'ils ont exprimé lors des entretiens collectifs.

Observation 7 : Les joueurs individualistes ont tendance à mentir sur les cartes conservées lorsque celles-ci sont individualistes (cf. Figure 47). Par exemple, après avoir conservé la conque, un joueur individualiste a dit *« je vous jure que ce sera utile pour le collectif »*, et un autre a dévoilé une autre carte à la place de la carte individualiste piochée. Cet élément pourra être ajouté comme cue dans l'arbre de détection d'informations critiques de Postdare.

Observation 8 : Nous avons observé un comportement de sabotage de la part d'un joueur individualiste suite à la décision des autres de l'éliminer. Il a fait une action inutile pour le collectif (chercher 1 bois alors qu'il y avait déjà 3 radeaux) et planifiait de transformer de la nourriture en eau avec le moulin à légumes afin d'engendrer une pénurie de nourriture.

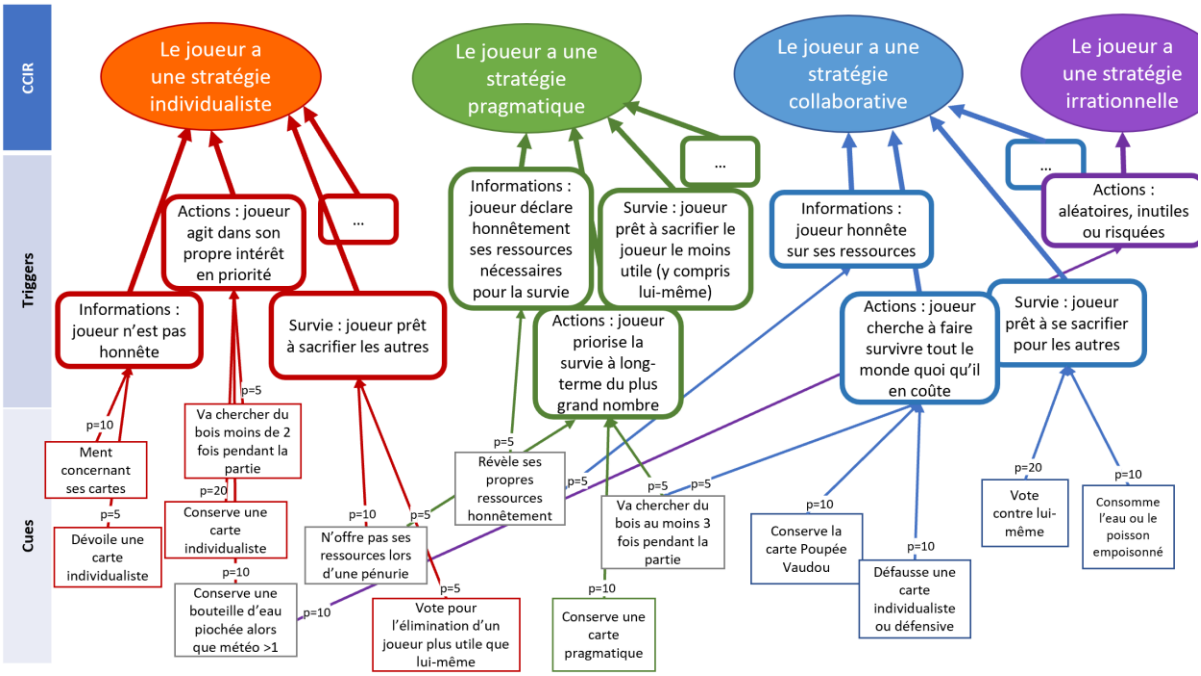

Figure 50 : Extrait de l'arbre de détection d'informations critiques de type ANTICIPE obtenu à l'issue des phases A et B de l'expérimentation

Suite aux hypothèses infirmées ou confirmées et aux nouvelles observations faites au cours de la phase B, le nouvel arbre de détection d'informations critiques comporte 120 cues. 11 cues identifiées lors de la phase A ont été supprimées car elles découlaient d'hypothèses qui ont été invalidées. D'autres cues ont

été reformulées pour être les plus objectives et factuelles possibles. Un extrait du résultat est présenté en Figure 50.

4.3.2.3 Modèle de prédiction des stratégies à partir des réponses aux questionnaires

Nous avons construit une ébauche de modèle de prédiction des stratégies à partir des réponses aux questionnaires Big-Five, MCQ-30, GDMS et MIST. Ce modèle s'appuie sur une élimination récursive des variables et une régression linéaire. Les variables retenues sont *« Conscience, contrôle, contrainte »* (issue du Big Five), *« Rationnel »* et *« Dépendant »* (issus du GDMS), *« Cognitive confidence »* et *« Cognitive self-consciousness »* (issues du MCQ-30).

Il avait été initialement prévu d'intégrer ces variables de prédiction en tant que *cues* dans une catégorie spéciale de l'arbre de détection d'informations critiques de Postdare, afin de donner un indice initial sur la stratégie la plus probable du joueur avant même qu'il remplisse le questionnaire de la main initiale idéale. Cependant, nous estimons que le faible nombre d'observations, et en particulier de stratégies individualistes, ne nous permet pas de valider ce modèle pour une utilisation au cours de la phase C de l'expérimentation.

Le modèle sera donc revu en intégrant les données récoltées au cours de la phase C (voir paragraphe *4.3.3.6 Révision du modèle de prédiction de la stratégie à partir du profil psychologique, métacognitif et décisionnaire*).

4.3.2.4 Limites de la phase B

Certaines limitations de cette phase de l'expérimentation méritent d'être soulignées.

Notamment, les joueurs ont été entraînés au jeu de Galèrapagos ensemble, au cours d'un tournoi où ils avaient pour consigne de jouer en collectif, une partie des points pour le tournoi étant attribués à la fin de la partie par les autres joueurs en fonction de leur collaboration. Ainsi, pour beaucoup, ils connaissaient déjà le style de jeu des autres et se sont influencés les uns les autres vers des comportements plus collaboratifs. Cela peut expliquer en partie le fait que nous avons observé peu de comportements individualistes au cours de la phase B. Les quelques joueurs plus « individualistes » au cours de cette phase ont déclaré que les consignes à orientation individualiste énoncées au début de l'expérimentation (ne pas montrer ses cartes, objectif de quitter l'île soi-même en priorité dans le groupe) incitaient à adopter des stratégies plus agressives comparées au tournoi, de même que les cartes agressives qui étaient en main dès le début du jeu.

Par ailleurs, l'ordre de distribution des cartes pendant la partie pourrait être amélioré. Les cartes « hache » et « canne à pêche » interviennent trop tôt : le rôle d'aller chercher du bois est ainsi « assigné » dès le premier tour de jeu. Nous n'avons donc pas pu observer le comportement des joueurs en l'absence de cette assignation.

4.3.3 Résultats de la phase C

4.3.3.1 Caractéristiques des participants

Le recrutement pour la phase C de cette expérimentation s'est fait auprès d'associations de jeux de société de la région bordelaise et auprès de collaborateurs THALES.

Cette phase s'est déroulée avec 22 participants âgés de 18 à 64 ans, dont 9 femmes et 13 hommes. 15 participants ont 32 ans ou moins (Figure 51).

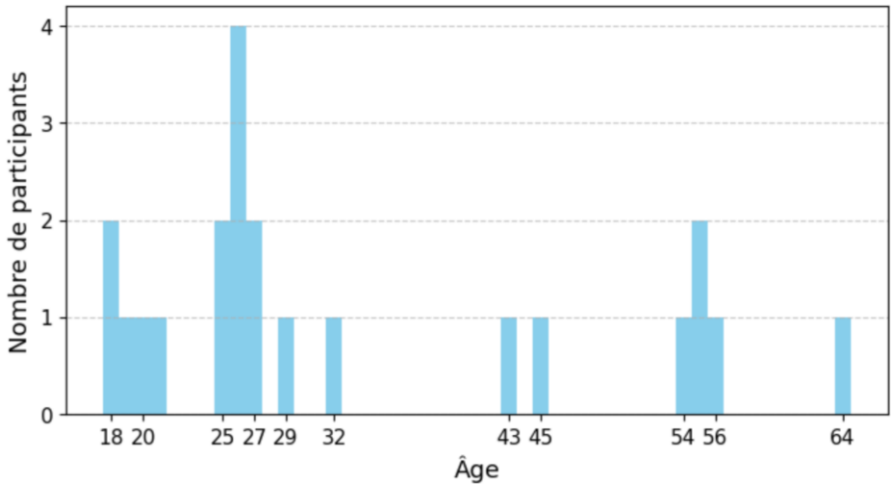

Figure 51 : Répartition des âges des participants de la phase C

Seul un participant n'a pas de diplôme, et la moitié a un niveau Bac+5 (Figure 52). 20 sont de nationalité française, et tous parlent couramment le français.

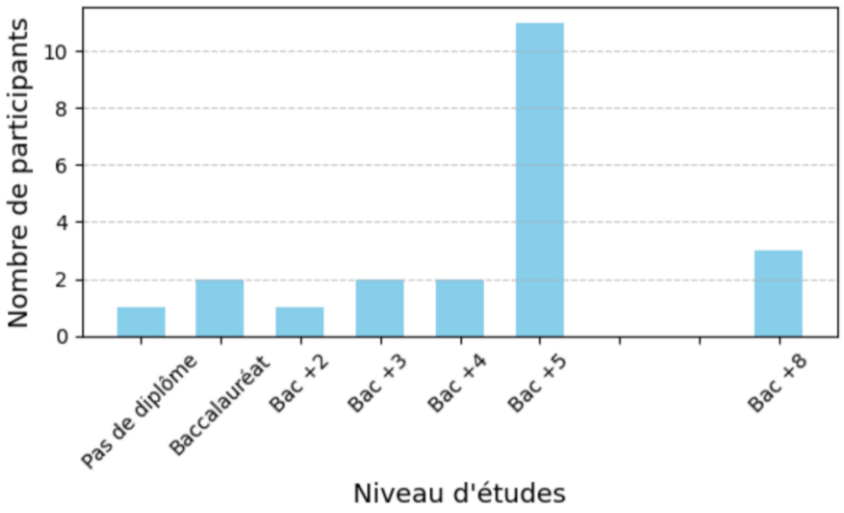

Figure 52 : Répartition des niveaux d'études des participants de la phase C

Le score MIST moyen est de 15,14/20, légèrement plus élevé que celui de la population évaluée par l'Université de Cambridge qui a conçu ce test (Maertens et al., 2023). Ce résultat indique que les participants ont tendance à être plutôt sceptiques que naïfs, présentant plus de faux négatifs que de faux positifs (Figure 53). En effet, le nombre moyen de faux négatifs est de 3,4, alors que le nombre moyen de faux positifs est de 1,5.

La moitié des passations s'est déroulée à distance, avec un partage d'écran permettant au participant de prendre le contrôle de la machine d'expérimentation sur laquelle se déroulaient les parties de Galèrapagos.

9 participants n'avaient jamais joué au Galèrapagos, et seulement 3 étaient experts du jeu. Cependant, 18 participants jouaient au moins une fois par mois à des jeux de société, ce qui leur a permis de prendre en main rapidement ce nouveau jeu dont les règles sont accessibles.

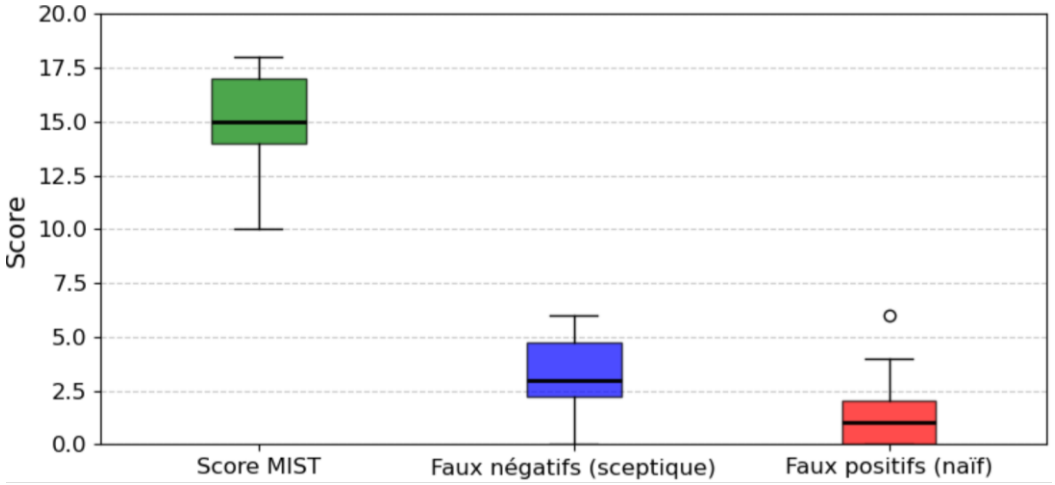

Figure 53 : Moyenne et écart-type des scores MIST des participants de la phase C incluant les faux négatifs (score « sceptique ») et les faux positifs (score « naïf »)

4.3.3.2 Statistiques des parties jouées

Au cours des 66 parties jouées pendant la phase C (3 pour chacun des 22 participants), 59 ont été gagnées par au moins 1 joueur, dont 46 gagnées par le participant. Ce résultat correspond aux consignes données à l'expérimentateur *agent* qui simulait les joueurs *Léa* et *Thomas*, et qui devait favoriser la survie du véritable participant le plus longtemps possible afin d'observer ses choix tout au long du jeu. Cette intervention passait par des ajustements discrets : une manipulation ponctuelle des cartes de *Léa* et *Thomas* possible uniquement avec l'interface expérimentateur (par exemple, pour obtenir une ressource manquante à la fin d'un tour), ou encore un seul vote de *Léa* ou *Thomas* contre le participant, étant donné qu'en cas d'égalité des votes, l'arbitrage automatique du logiciel est toujours en faveur du joueur réel.

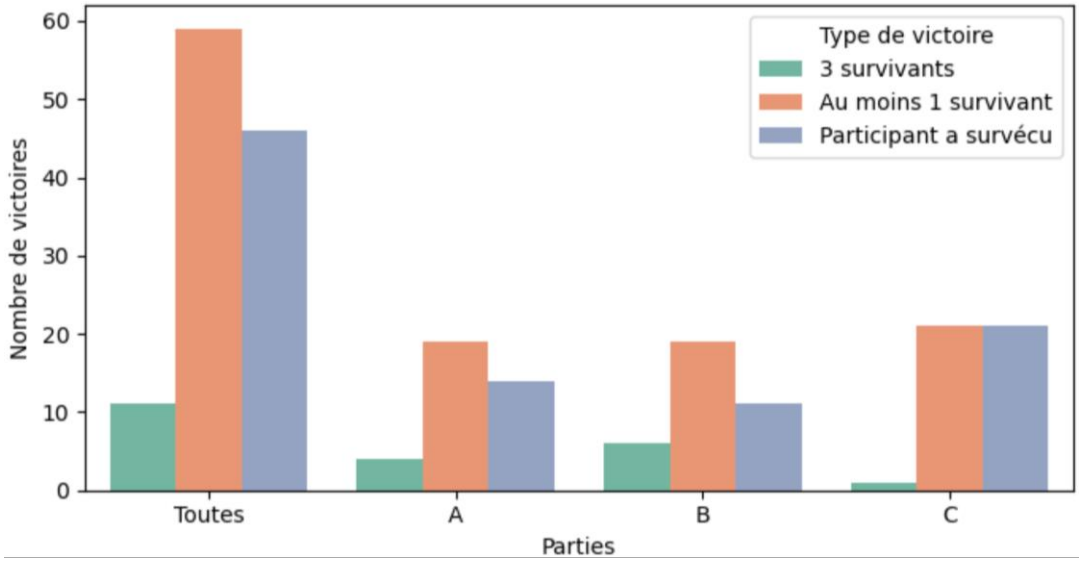

Figure 54 : Nombre de parties gagnées par le participant ou par le groupe au cours de la phase C, pour toutes les parties réunies et pour les parties A, B et C séparément

Le taux de réussite du participant est donc d'environ 70%. Cependant, seulement 11 parties ont été gagnées par les 3 joueurs, soit un taux de victoire collective de 17%. Cela correspond également à notre volonté de rendre la victoire à 3 joueurs relativement difficile, afin de forcer le participant à faire des sacrifices et à faire face à de véritables choix stratégiques.

Les parties A et B présentent des statistiques relativement similaires. Pour la partie C, nous observons un taux de réussite du participant plus élevé : en effet, elle a été gagnée par le participant 21 fois sur 22, ce qui signifie qu'un seul participant a échoué. À l'inverse, les parties A et B ont été gagnées respectivement 14 et 11 fois par le participant. En revanche, la partie C n'a vu les 3 joueurs survivre collectivement qu'une seule fois au cours de l'expérimentation, contre 4 et 6 fois pour les parties A et B (Figure 54).

4.3.3.3 Interprétation de la stratégie via Postdare

Au cours des 66 parties jouées, nous comptons en moyenne 16,4 cues par partie (écart-type 4,6) qui ont été activées dans Postdare, dont 12,6 en moyenne (écart-type 4) sélectionnées automatiquement et 3,8 en moyenne (écart-type 2,1) sélectionnées manuellement (Figure 55).

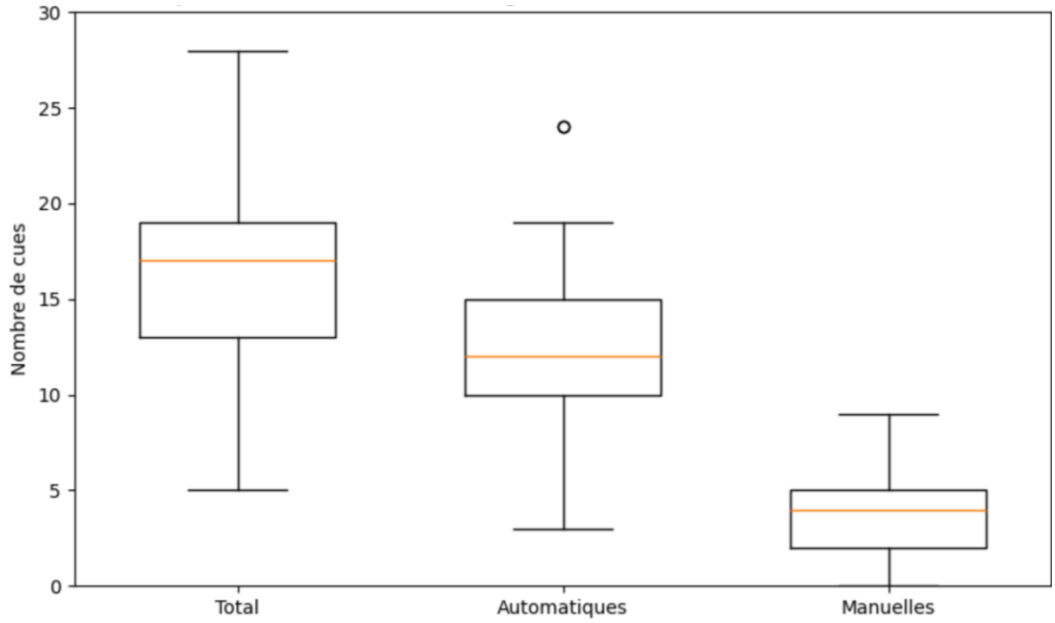

Figure 55 : Répartition des différentes catégories de cues sélectionnées sur Postdare au cours des parties jouées pendant la phase C

Nous avons analysé la stratégie déclarée par le joueur au cours 1) de son questionnaire de ressenti de la partie (par exemple, *« j'avais pour objectif de faire survivre tout le monde »* a été interprété comme une stratégie collaborative) et 2) de l'entretien final. Celles-ci diffèrent parfois (différence importante pour 3 parties sur un total de 66). Les différences observées sont généralement dues aux variations de stratégie au cours d'une même partie (un joueur peut avoir un objectif collaboratif en début de partie puis adopter un comportement plus individualiste en fonction de la tournure des événements). Mais elles peuvent aussi être le résultat d'une reconstruction inadaptée des événements dans la mémoire du participant ou à un biais de désirabilité sociale. Par exemple, un joueur qui tend à être individualiste pourrait être réticent à l'admettre.

Une analyse qualitative des stratégies au cours des parties jouées comparée à la détection par Postdare a révélé que cette dernière est généralement correcte. En effet, lorsque les cas où le joueur ne déclare pas la même stratégie pendant le questionnaire de ressenti de la partie et l'entretien sont considérés comme des erreurs pour Postdare, la détection reste correcte dans 65% des cas (pire cas considéré – voir Tableau 22). La stratégie détectée par Postdare est la même que celle déclarée par le joueur pendant l'entretien dans 88% des cas. Nous nous sommes intéressés aux erreurs entre des stratégies opposées, c'est-à-dire lorsque Postdare détecte une stratégie totalement opposée à celle déclarée par le joueur

(individualiste vs. collaboratif) : dans le pire des cas à nouveau, on observe 9 erreurs importantes sur les 66 parties jouées (Tableau 22 – dernière ligne).

Tableau 22 : Correspondances entre la stratégie dominante évaluée par Postdare, le questionnaire de ressenti de la partie et l'entretien

Stratégie dominante évaluée par :	Nb /66	%
Postdare / Questionnaire de ressenti (Qr)	48 corrects	73%
Postdare / Entretien (E)	58 corrects	88%
Postdare / (Qr et E)	43 corrects	65%
Confusions individualiste / collaboratif (QR ou E)	9 erreurs	14%

La détection est moins précise dans le cas de certains joueurs expérimentés au Galèrapagos. En effet, ces joueurs tendent à avoir un comportement collaboratif en apparence tout en préservant secrètement la possibilité de jouer de manière individualiste pour survivre eux-mêmes si les ressources ne permettent pas une victoire collective. Cela s'est traduit notamment au cours des 3 parties (sur les 66 au total) jouées par des joueurs expérimentés, pour lesquelles la stratégie détectée par Postdare ne correspond ni à la stratégie déclarée dans le questionnaire de ressenti de la partie, ni au cours de l'entretien. Dans les trois cas, les pourcentages des stratégies individualiste, pragmatique et collaborative détectées par Postdare sont relativement proches, ce qui correspond à des changements de stratégie au cours du jeu. Par exemple, pour la deuxième partie concernée, le participant a déclaré son objectif final individualiste dans le questionnaire et au cours de l'entretien. Cependant, il est resté collaboratif pendant toute la partie ; la stratégie dominante détectée par Postdare est donc « collaboratif ». Néanmoins, nous observons une transition vers une stratégie plus individualiste à la fin de la partie. Ainsi, pour bien comprendre l'évolution de la stratégie au cours du jeu, il faut consulter les stratégies détectées par Postdare en fonction des phases du jeu ; se contenter du résultat final ne permet pas toujours de dégager du sens de la stratégie observée, à cause de la nature du jeu qui appelle des changements de stratégie dans certaines situations.

4.3.3.4 Analyse des stratégies adoptées

Pour 54 parties, la stratégie dominante du participant détectée via Postdare était collaborative, elle était individualiste pour 9 parties et pragmatique pour 3 parties.

Les participants ont été légèrement plus collaboratifs au cours de la partie B (Figure 56). En revanche, il ne semble pas y avoir d'effet de l'ordre des parties sur la stratégie dominante des joueurs. Aucun participant n'a été catégorisé comme étant majoritairement irrationnel, même si certains ont exprimé plus de comportement irrationnel que d'autre (jusqu'à 29% détectés par Postdare au cours d'une partie).

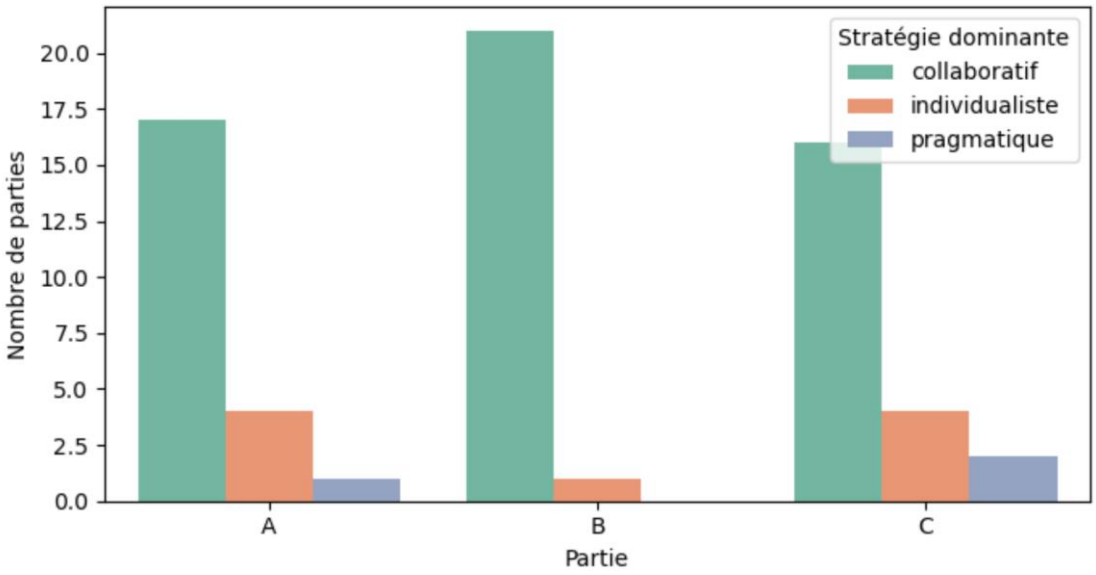

Figure 56 : Stratégies dominantes des participants pour chaque partie possible

Nous ne distinguons pas dans les données de différence majeure entre les stratégies des joueurs qui jouent fréquemment à des jeux de société et les joueurs qui y jouent moins fréquemment, mais il semble que plus de joueurs expérimentés au Galèrapagos adoptent des stratégies individualistes (Figure 57). Cependant, nous ne disposons pas de suffisamment de données pour confirmer statistiquement cette observation.

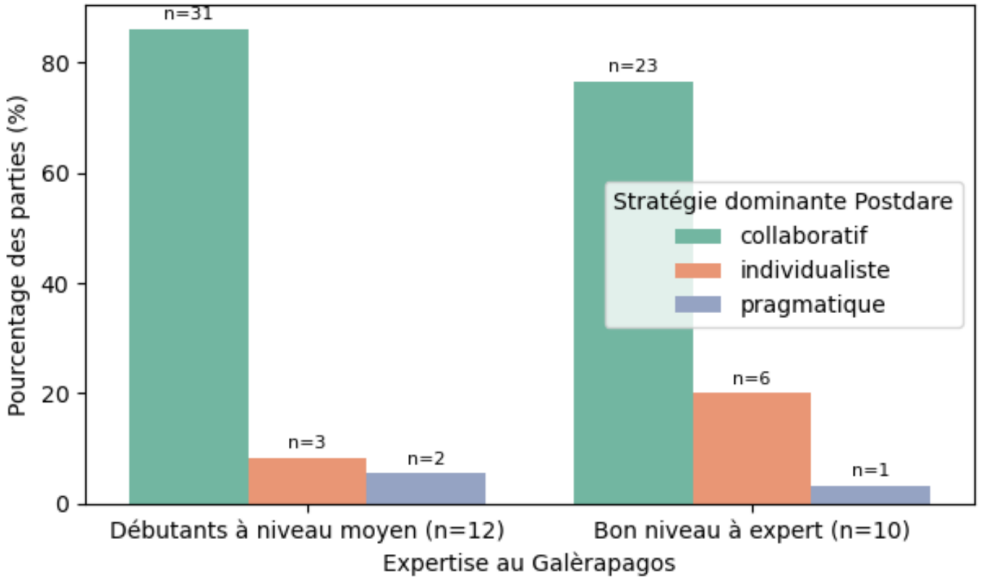

Figure 57 : Stratégies des joueurs en fonction de leur niveau d'expertise au Galèrapagos

4.3.3.5 Influence de la stratégie via les messages d'influence

Au cours des parties observées, sur les 22 participants, 18 étaient initialement collaboratifs et ont donc reçu des messages d'influence vers une stratégie plus individualiste, tandis que 4 seulement ont reçu des messages d'influence vers une stratégie collaborative.

Évolution des tendances stratégiques d'après les mains initiales idéales :

Dans les réponses aux questionnaires de main initiale idéale (complétés avant la première partie observée, avant la deuxième partie observée et donc avant les tentatives d'influence, et après la troisième partie observée donc après les tentatives d'influence), nous avons interprété une tendance stratégique pour chaque main choisie en fonction des usages des différentes cartes sélectionnées par le joueur. Nous remarquons ici que la grande majorité des joueurs a une tendance stratégique initiale collaborative à partiellement pragmatique et partiellement collaborative.

La tendance stratégique semble stable pour presque tous les joueurs, sauf deux participants pour lesquels il existe une différence notable sur le dernier questionnaire complété : un participant (n°6) est passé d'une main initiale idéale collaborative à une main individualiste (en bleu sur la Figure 58), et un autre (n°15) est passé d'une main entre pragmatique et individualiste à une main entre pragmatique et collaborative (en vert clair sur la Figure 58). Ces différences vont toutes deux dans le sens de l'influence à laquelle nous les avons soumis au cours des parties 2 et 3.

Pour 4 participants (n°6, 7, 9 et 19), la main initiale idéale diffère de leur objectif déclaré lors des questionnaires de ressenti de la partie et lors de l'entretien : pour chacun d'entre eux, la main initiale idéale correspond à leur tendance stratégique initiale, tandis que lors du questionnaire de ressenti de la partie et lors de l'entretien, ils ont déclaré une tendance inverse, qui correspondait au sens de l'influence à laquelle ils ont été soumis.

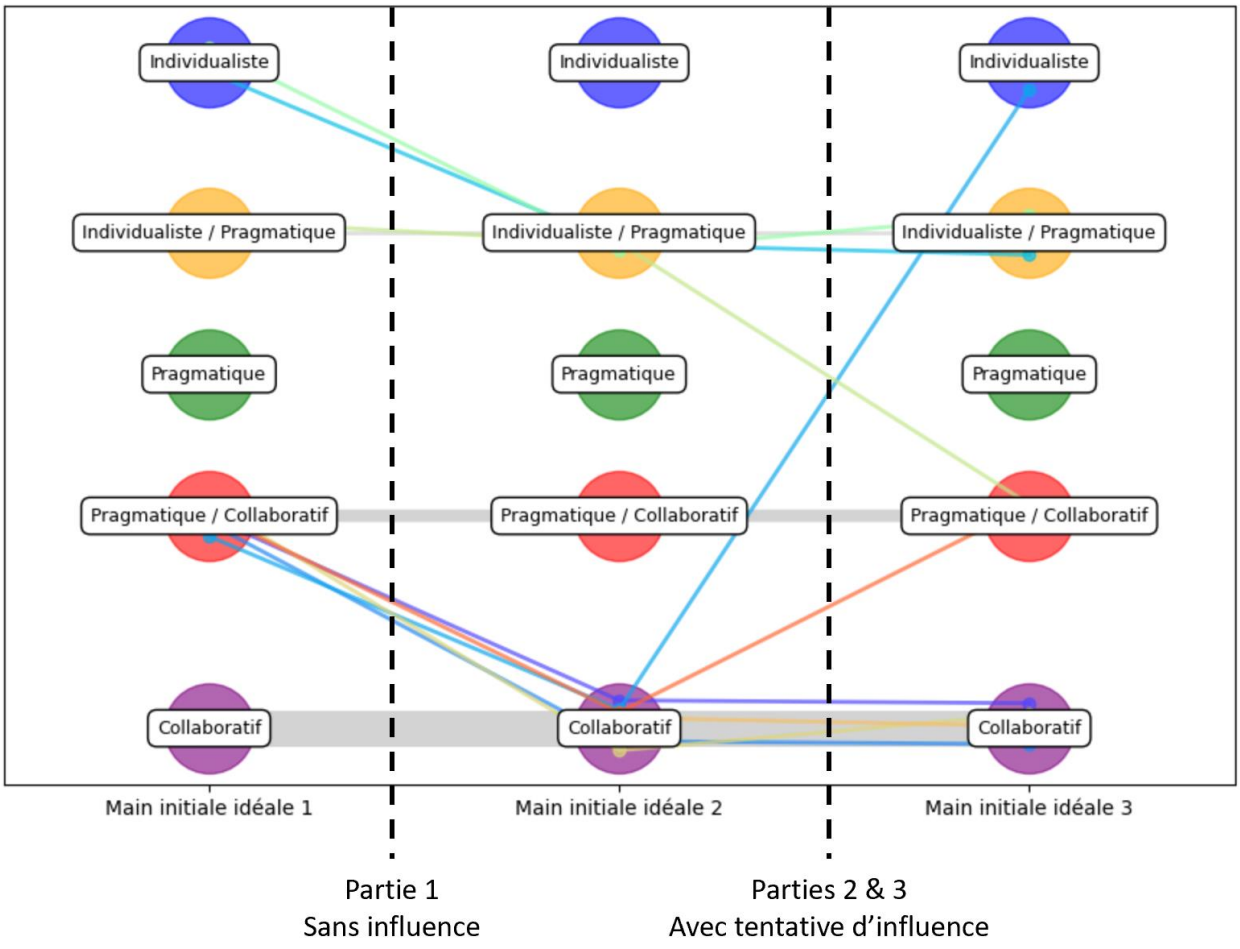

Figure 58 : Évolution des tendances stratégiques des participants de la phase C d'après leurs mains initiales idéales

Évolution de la stratégie dominante d'après Postdare :

La Figure 59 montre qu'il existe plus de modifications importantes de la stratégie au cours des parties jouées que dans le questionnaire de la main initiale idéale (Figure 58) : celui-ci semble donc mieux refléter les tendances stratégiques à long-terme, tandis que la stratégie observée au cours des parties jouées reflète plus les conditions de la partie et l'influence immédiate. Nous constatons que les 3 joueurs (représentés par des traits bleus) adoptant une stratégie dominante individualiste en début d'expérimentation ont évolué vers une stratégie plus collaborative, dès la deuxième partie jouée pour deux d'entre eux. Au contraire, 4 joueurs initialement collaboratifs et que nous avons tenté d'influencer vers une stratégie individualiste (traits rouges), ont effectivement adopté une stratégie plus individualiste, dont l'un d'entre eux seulement pour la deuxième partie, et les trois autres seulement pour la troisième partie jouée.

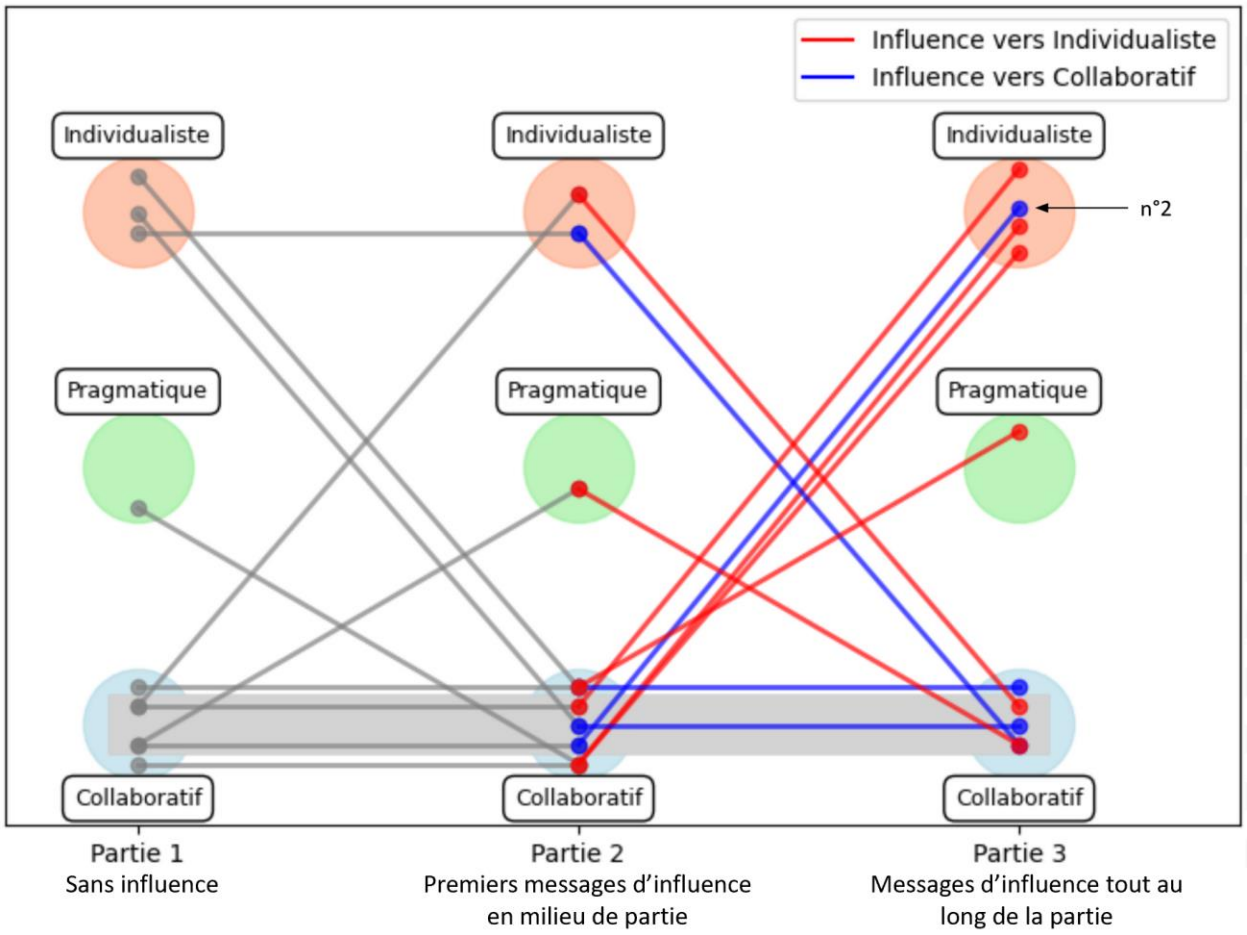

Figure 59 : Évolution de la stratégie dominante des participants de la phase C au cours des parties jouées d'après Postdare

La Figure 59 permet d'observer qu'un participant (joueur n°2, représenté par un trait bleu allant de « Collaboratif » à la partie 2 à « Individualiste » à la partie 3) à la stratégie dominante initialement collaborative est devenu plus individualiste malgré que nous ayons tenté de l'influencer vers une stratégie collaborative. Pour ce participant, nous avons observé sur Postdare un taux important de stratégie individualiste (même si ce n'était pas la stratégie dominante), et en analysant ses réponses aux différents questionnaires, nous avons estimé que sa tendance générale était plutôt à l'individualisme, c'est pourquoi nous avons choisi de l'orienter vers une stratégie plus collaborative. Ainsi, nous constatons qu'il est insuffisant de s'intéresser uniquement à la stratégie dominante relevée par Postdare, il ne faut pas négliger les variations stratégiques naturelles du joueur.

Évolution de la stratégie des joueurs soumis à une influence « vers collaboratif » :

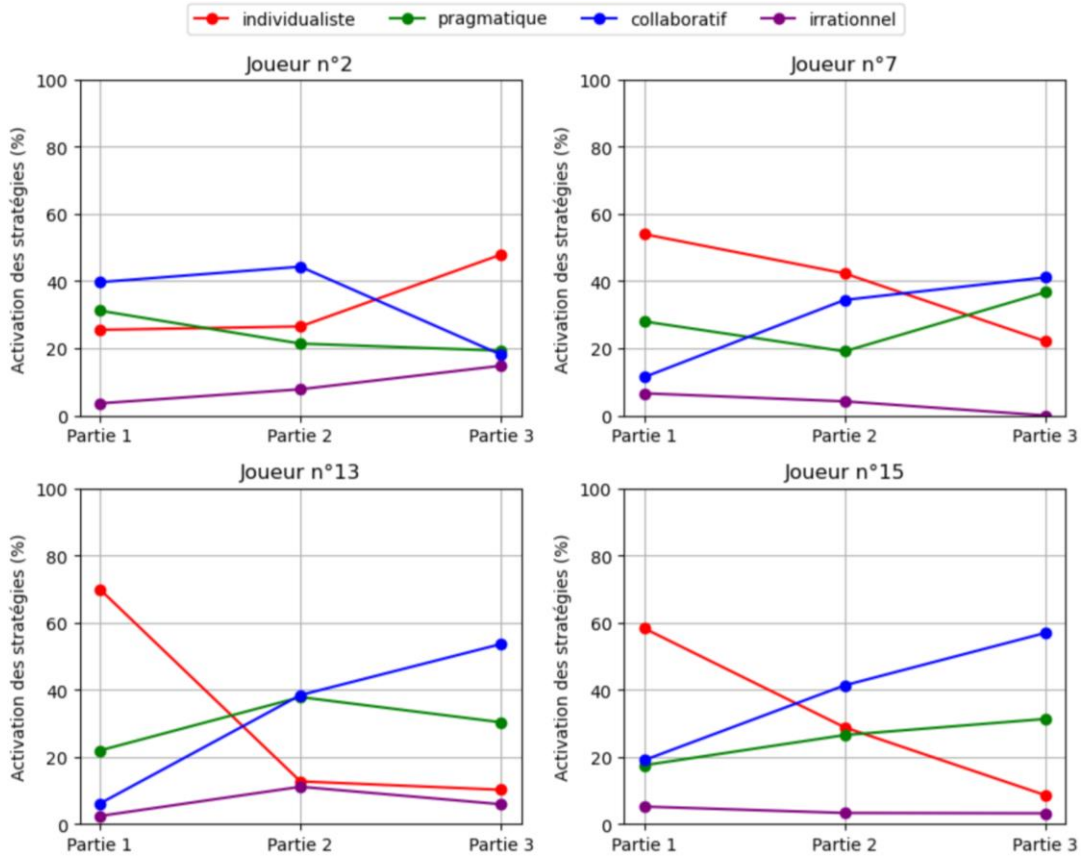

Figure 60 : Évolution de l'activation des différentes stratégies sur Postdare pour les 4 participants de la phase C influencés vers une stratégie collaborative

La Figure 60 présente, d'après Postdare, l'évolution de l'activation de chaque stratégie pour les 4 joueurs qui ont été orientés vers une stratégie plus collaborative. Le premier graphique obtenu (en haut à gauche) présente les résultats du participant n°2 mentionné précédemment, pour lequel la stratégie initiale a été identifiée comme à tendance individualiste malgré une évaluation plus collaborative d'après Postdare. Pour les trois autres participants, nous observons que l'activation de stratégie individualiste a diminué au cours des trois parties, tandis que l'activation de stratégie collaborative a augmenté tout au long de l'expérimentation.

Évolution de la stratégie des joueurs soumis à une influence « vers individualiste » :

Les participants que nous avons tenté d'influencer vers une stratégie individualiste étant nombreux (18), nous avons choisi de représenter sur la Figure 61 des clusters des graphiques ayant la forme la plus proche. Nous observons que les clusters 1 et 4 incluent 5 passations pour lesquelles le taux d'attitudes collaboratives a effectivement diminué, tandis que le taux d'individualisme a augmenté (en particulier pour le cluster 1). Pour les clusters 2 et 3, nous ne distinguons pas de changement majeur, voire une augmentation de la collaborativité, qui va donc à l'opposé de l'influence à laquelle les joueurs concernés ont été soumis. Au total, 8 participants influencés vers une stratégie individualiste voient leur taux d'individualisme augmenter, leur taux de collaborativité diminuer, ou les deux, d'au moins 15% au cours de la partie 2 ou 3 (ou des deux). De même, 5 d'entre eux ont vu leur stratégie évoluer d'au moins 25% dans le sens vers lequel ils ont été influencés.

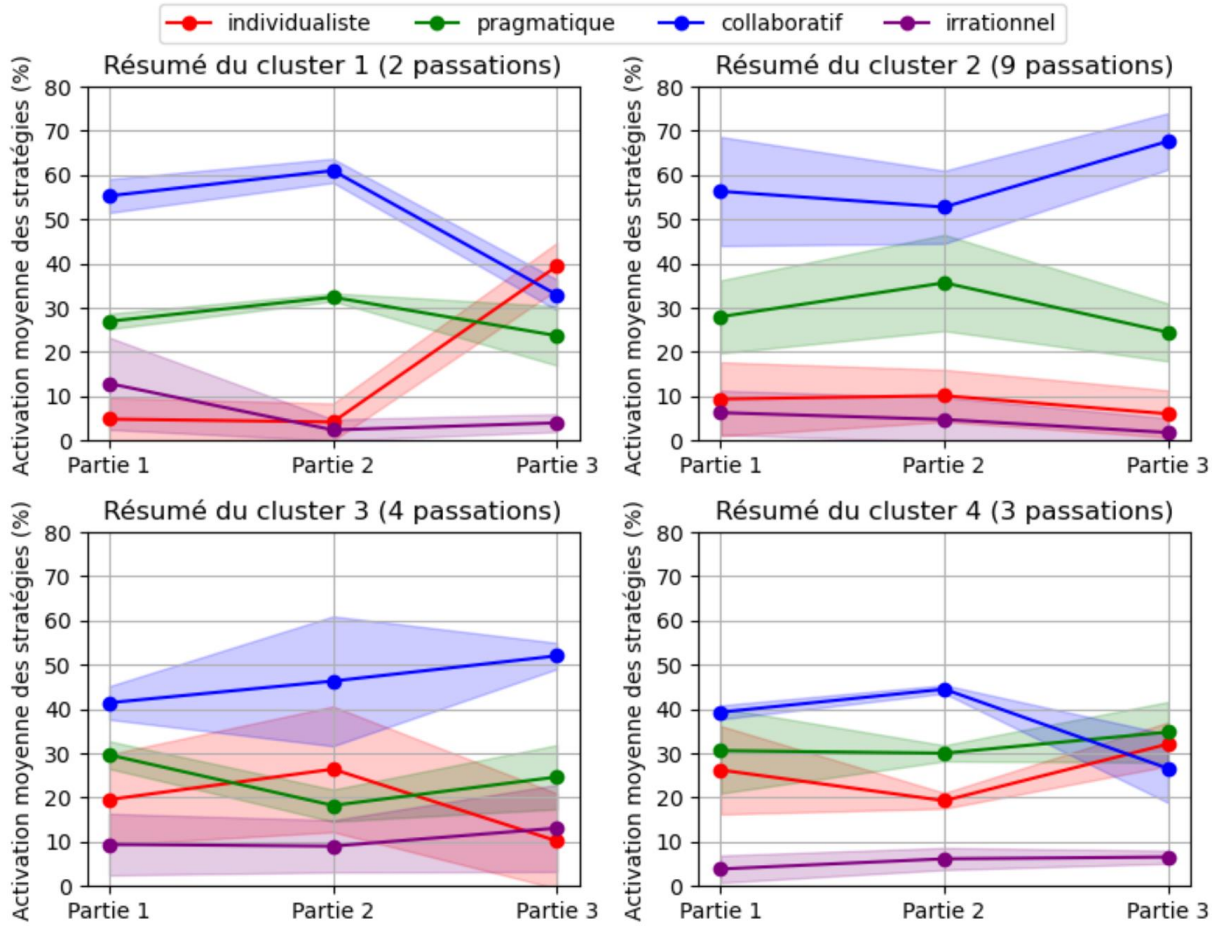

Figure 61 : Évolution de l'activation des différentes stratégies sur Postdare pour les 18 participants de la phase C soumis à une influence vers une stratégie individualiste, représentés sous forme de 4 clusters

Évolution de la stratégie des joueurs d'après l'analyse des entretiens :

Dans les deux paragraphes précédents, nous avons établi que 8 participants sur les 22, soit 36%, ont vu leur stratégie évoluer considérablement (>25%) pendant la partie 2, la partie 3, ou les deux, dans le sens vers lequel ils ont été influencés. Nous pouvons nous demander si ces changements de stratégie sont dus à l'influence ou à d'autres facteurs tels que le déroulement des parties, le comportement des autres joueurs ou des choix personnels.

Comme nous l'avons vu dans le paragraphe *4.2.2.3 Phase C : Détecter la stratégie en temps réel et tenter de l'influencer*, le comportement des joueurs fictifs (Léa et Thomas) était contrôlé afin que Léa ait une tendance plutôt collaborative et Thomas une tendance plutôt individualiste, afin de permettre au participant de choisir sa propre stratégie. La plupart des participants ont effectivement perçu cette différence et nombre d'entre eux ont cherché à s'allier avec Léa à un moment ou un autre durant l'expérimentation, et parfois à éliminer Thomas. Cependant, certains participants ont indiqué qu'ils trouvaient Léa et Thomas plutôt collaboratifs, tout en reconnaissant un trait plus « provocateur » chez Thomas, ce qui peut s'expliquer par la consigne de l'expérimentateur de faire en sorte que le participant survive jusqu'à la fin de la partie afin d'observer ses choix pendant le scénario complet. Nous posons donc l'hypothèse que le comportement des joueurs fictifs ne devrait pas avoir d'influence majeure sur la stratégie des participants, ou une influence constante au cours des trois parties qui ne devrait pas interférer avec l'influence volontaire par les « protips ».

La Figure 56 (p. 148) a montré que les participants avaient été légèrement plus collaboratifs pendant la partie B. Nous observons trois parties de plus avec un participant à stratégie dominante collaborative pour cette partie. Or, nous constatons que deux des trois participants qui ont été orientés vers une stratégie plus collaborative et qui ont effectivement changé de stratégie avaient joué la partie B comme dernière partie, et nous observions déjà un changement de stratégie important pendant la deuxième partie jouée (Figure 60, les deux graphiques du bas). Cela peut donc expliquer la différence de répartition des stratégies entre les parties A, B et C, et nous pouvons supposer que la partie jouée n'a pas ou peu d'influence sur la stratégie adoptée.

Au cours des entretiens, diverses raisons expliquant les changements de stratégie ont été mises en valeur :

- Un joueur (n°6) initialement collaboratif a adopté une stratégie de plus en plus individualiste tout au long de l'expérimentation en se rendant compte qu'il était difficile de gagner à trois joueurs.
- Un joueur (n°7) initialement individualiste, expérimenté dans les jeux de société mais qui découvrait le Galèrapagos, a souhaité tester un style stratégique différent pour la dernière partie, mais pense avoir été également influencé par les messages d'influence qui l'incitaient à devenir plus collaboratif.
- Un joueur (n°15) initialement individualiste a adopté des stratégies de plus en plus collaboratives pendant les parties jouées, car après avoir réussi à survivre seul, il a souhaité tenter de survivre à deux lors de la deuxième partie puis à trois lors de la troisième afin d'élever le niveau de difficulté.
- Un joueur (n°19) initialement collaboratif est devenu plus individualiste au cours de la deuxième partie, à cause de l'individualisme de Thomas et car son manque d'expérience au Galèrapagos lui a fait penser que Léa avait aussi une tendance individualiste par incompréhension de certaines de ses actions.

Analyse de l'effet des messages d'influence d'après les entretiens :

D'après les entretiens, plusieurs joueurs ont trouvé qu'il y avait beaucoup d'informations et que les parties se déroulaient vite, ne leur permettant pas toujours de prendre en compte tout ce qu'il se passait. Une partie d'entre eux a donc ignoré les messages « protips » (messages d'influence) ou les a *« lus en diagonale »* parce qu'ils les trouvaient répétitifs ou trop longs et parfois agaçants, inintéressants ou en désaccord avec leur propre stratégie, ou encore parce qu'ils connaissaient déjà bien le jeu. Ainsi, 9 participants ont déclaré ne pas avoir lu la totalité des protips ou les avoir lus sans y prêter beaucoup d'attention. La plupart d'entre eux (8/9) disent ne pas avoir été influencés par ces mêmes messages. Parmi les participants qui ont déclaré les avoir lus avec attention, 4 estiment avoir été influencés et avoir modifié leur stratégie après avoir lu certains messages. Si ces chiffres peuvent nous donner des indices sur l'efficacité des tentatives d'influence, il est à noter que les participants n'ont pas une vision objective sur leur propre stratégie et leur résistance à la manipulation. En effet, la psychologie sociale a montré que les raisons que les individus donnent à leurs comportements sont souvent des reconstructions *a posteriori* et non des causes réelles (Nisbett & Wilson, 1977). Nous pouvons donc estimer que les participants ne savent pas toujours ce qui les a vraiment influencés, mais tentent de le rationaliser *a posteriori*.

Le Tableau 23 montre que les messages d'influence jugés utiles et intéressants par le plus de participants sont les deux qui correspondent à des astuces sans intention d'influence (jugés intéressants par 12 et 10 personnes respectivement). Les 4 participants qui ont été soumis à l'influence vers une stratégie collaborative ont tous trouvé intéressant le message d'influence incluant des exemples et détails (message 7 – cf. Tableau 21), celui-ci semble donc avoir été très efficace. Parmi les 18 participants soumis à l'influence vers une stratégie individualiste, 4 ont accordé un intérêt particulier à celui présentant des

répétitions (message 12) et 3 ont accordé un intérêt à celui présentant des exemples et détails (message 11).

Tableau 23 : Intérêt manifesté par les participants pour les différents « protips » présentés

Protip	Fréquence à laquelle il a été trouvé intéressant	Fréquence d'apparition totale	% d'intérêt des participants
2 – Astuce	12	22	55%
1 – Astuce	10	22	45%
7 – Exemples & détails (vers collaboratif)	4	4	100%
12 – Répétitions (vers individualiste)	4	18	22%
11 – Exemples & détails (vers individualiste)	3	18	17%
10 – Sources (vers individualiste)	2	18	11%
6 – Sources (vers collaboratif)	1	4	25%
4 – Publicité	1	22	5%

L'efficacité des messages d'influence est donc mitigée, ce qui est notamment dû au format sous lequel ils ont été présentés et qui a conduit une partie des participants à les ignorer. Ils pourraient être améliorés avec un format différent qui les mette plus en valeur, et un contenu plus varié afin de masquer la répétition de l'injonction à jouer de manière plus collaborative ou plus individualiste.

4.3.3.6 Révision du modèle de prédiction de la stratégie à partir du profil psychologique, métacognitif et décisionnaire

À l'issue de la phase B de l'expérimentation, un modèle de prédiction de la stratégie basé sur une régression linéaire a été développé à partir des réponses aux questionnaires MIST, métacognitif, de personnalité et de style de prise de décision (paragraphe *4.3.2.3 Modèle de prédiction des stratégies à partir des réponses aux questionnaires*).

Pour tester ce modèle avec les données de la phase C, nous avons utilisé la stratégie déduite du premier questionnaire de main initiale idéale. En effet, nous avons constaté que ce questionnaire permettait d'évaluer la tendance stratégique des joueurs de manière plus stable que la simple observation des stratégies mises en œuvre au cours des parties jouées. De plus, le premier questionnaire de la main initiale idéale est complété avant les parties jouées, et donc avant toute influence possible des messages d'influence ou des autres joueurs. D'après ce modèle, les variables les plus significatives proviennent du questionnaire Big Five et du GDMS : conscience-contrôle-contrainte (niveau d'organisation, discipline et fiabilité), agréabilité-altruisme-affection (tendance à être compatissant, coopératif et bienveillant) et un style de prise de décision rationnel (approche logique, structurée et basée sur les faits). Cependant, testé avec les données de la phase C, le modèle ébauché utilisant une régression linéaire n'est pas concluant. En effet, en combinant les données des phases B et C, son R carré ajusté qui était de 0,82 avec les données de la phase B chute à 0,36 (Tableau 24). De même, la racine carrée de l'erreur quadratique moyenne (RMSE) augmente en combinant les données des phases B et C, montrant que les prédictions sont alors éloignées des valeurs réelles. Il semblerait qu'avec les données de la phase B uniquement, le modèle présentait un surapprentissage important et n'était en réalité pas généralisable.

Tableau 24 : Indicateurs de performance du modèle de régression linéaire

Phase	R^2	R^2 ajusté	RMSE
Régression sur phase B (n=12)	0,902	0,822	0,24
Régression sur phase C (n=22)	0,622	0,503	0,39
Régression sur phases B&C (n=34)	0,453	0,355	0,51

La Figure 62 montre que les résidus ne semblent pas être aléatoires et suggèrent une relation non-linéaire entre les variables explicatives et la variable à prédire (soit la stratégie).

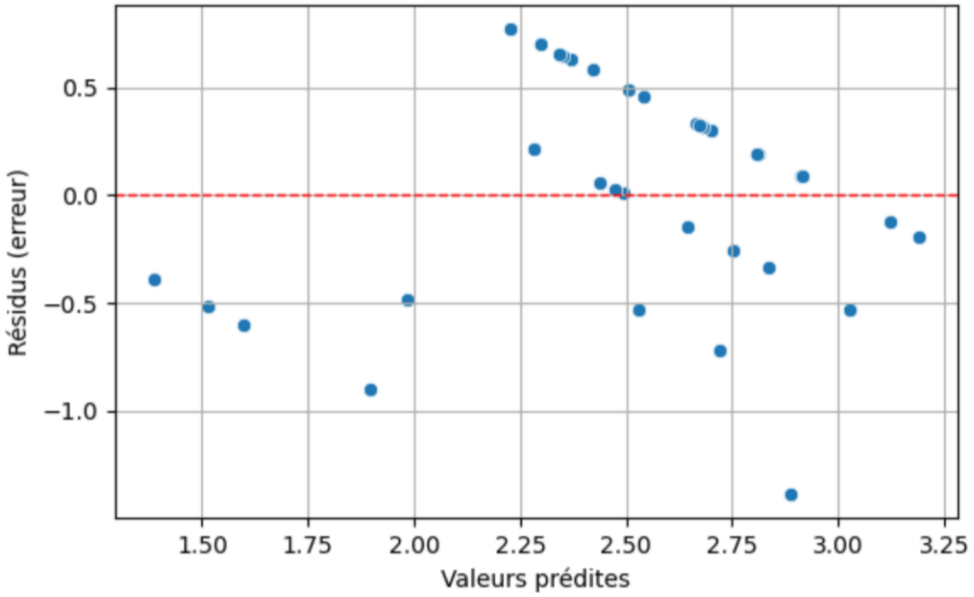

Figure 62 : Graphique des résidus du modèle de régression linéaire (phases B&C, n=34)

Nous avons donc tenté d'appliquer d'autres modèles. Par exemple, une régression par Random Forest avec et sans élimination des 7 variables les moins corrélées à la stratégie. Ce modèle produit comme variables les plus explicatives conscience-contrôle-contrainte (Big Five) et un style de prise de décision rationnel, mais aussi émotions négatives-névrosisme-nervosité (Big Five). Cependant, ce modèle n'est à nouveau pas généralisable.

Tableau 25 : Rapport de classification du modèle de Random Forest Classifier sur le jeu de données de test

Classe à prédire	Precision	Recall	F1-score	Support
0 (stratégie individualiste à pragmatique)	0,64	0,56	0,6	16
1 (stratégie collaborative)	0,65	0,72	0,68	18

Tableau 26 : Matrice de confusion du jeu de données de test pour le modèle de Random Forest Classifier

		Prédit	
		0 (indiv. à prag.)	**1** (collab.)
Réel	**0** (indiv. à prag.)	9	7
	1 (collab.)	5	13

Le meilleur modèle obtenu est basé sur un classifieur de type Random Forest avec validation croisée 5-fold, tentant de prédire 2 classes de stratégie rassemblant dans une classe les stratégies individualistes à pragmatiques, et dans l'autre les stratégies collaboratives (Tableau 25). Ce modèle semble discriminer les 2 classes de stratégies, avec une accuracy moyenne d'environ 64% en validation croisée (écart-type 12,8%). La matrice de confusion (Tableau 26) montre que le modèle reconnaît correctement la majorité des stratégies collaboratives (recall = 0,72), mais confond encore certaines stratégies individualistes ou pragmatiques avec des comportements plus collaboratifs (recall = 0,56). Les variables les plus contributives à la prédiction pour ce modèle sont un style décisionnel spontané (GDMS), la présence de croyances positives (MCQ-30) et émotions négatives-névrosisme-nervosité (Big Five).

Nous concluons de ces résultats qu'un classifieur de type Random Forest semble relativement efficace pour prédire si le joueur est collaboratif ou non, mais ne permet pas de prédire les tendances stratégiques de manière fine. La faible taille du jeu de données (34 observations en rassemblant les phases B et C de l'expérimentation) ne permet pas d'obtenir des modèles très robustes dans tous les cas. De la même façon, le score MIST (sensibilité à la désinformation) n'a été retenu par aucun modèle comme étant un prédicteur efficace de la stratégie.

4.4 Conclusion de l'expérimentation 2

La méthodologie d'arbre de détection d'informations critiques du système ANTICIPE permet-elle d'évaluer la stratégie du joueur en temps-réel ? À cette question de recherche, nous pouvons répondre que le système ANTICIPE (via l'outil Postdare) permet effectivement de détecter efficacement la stratégie d'un joueur en temps réel au cours d'une partie. Le questionnaire de la main initiale idéale, quant à lui, permet d'évaluer les tendances stratégiques du joueur à plus long terme.

Peut-on mener une stratégie qui influence les décisions prises ou l'issue d'un jeu de stratégie (Galèrapagos), sans que la personne s'en aperçoive et en s'appuyant sur les signaux faibles indiquant sa stratégie et sur des éléments de sa cognition ? Les messages d'influence ont été efficaces dans certains cas et 4 joueurs sur les 22 participants ont déclaré avoir modifié leur stratégie en fonction de ceux-ci. Cependant, certains joueurs ont exprimé avoir été conscients de la tentative de manipulation, ce qui traduit un manque de finesse dans les messages et la manière dont ils ont été présentés, aussi bien qu'une méfiance introduite par le contexte expérimental. Parmi les messages destinés à influencer, les participants ont accordé un intérêt particulier à ceux présentant des exemples et des détails (surtout pour les participants initialement individualistes et orientés vers une stratégie collaborative) et ceux présentant des répétitions dans le message.

Le modèle de prédiction de la stratégie à partir des données des questionnaires a seulement retenu des données liées aux questionnaires GDMS, Big Five et MCQ-30, le questionnaire MIST sur la sensibilité à la désinformation ne semblant pas être un prédicteur de la stratégie. Parmi les modèles testés, seul un classifieur de type Random Forest présente des performances acceptables ; cependant, il permet seulement de discriminer les stratégies collaboratives des autres et ne produit pas une prédiction fine de la stratégie. Il est important de souligner que les questionnaires retenus ne suffisent pas à prédire le comportement d'une personne dans différentes situations. Il est donc nécessaire de rechercher d'autres signaux faibles pour évaluer automatiquement la stratégie avec plus de précision, d'où l'importance du travail effectué ici et du système Postdare (ou ANTICIPE), qui a le potentiel de rassembler de nombreuses sources de données de manière lisible.

Certaines limites de notre étude sont à souligner. Notamment, l'échantillon de cette expérimentation est majoritairement composé de jeunes adultes, les 18-30 ans représentant la majorité des participants. Ce biais d'âge limite la représentativité du panel, cette tranche n'étant généralement pas impliquée dans les décisions stratégiques susceptibles d'avoir un effet important sur d'autres personnes. Pour accroître la robustesse du modèle, il serait souhaitable d'élargir le recrutement à des profils plus diversifiés, incluant des décideurs expérimentés.

Comme précisé précédemment, Postdare a des difficultés pour identifier les joueurs individualistes qui tendent à vouloir paraître collaboratifs pour tromper les autres concernant leurs intentions : même s'il détecte de l'individualisme, il déclare tout de même la stratégie dominante comme étant collaborative. Une adaptation du poids des cues associées pourrait permettre d'obtenir une meilleure classification. Dans tous les cas, une analyse de l'évolution des styles stratégiques détectés par Postdare au cours de la partie en fonction du déroulement des phases de jeu permet de mieux comprendre l'état d'esprit du joueur. Un affichage comparatif de la stratégie dominante au cours de la partie vs. au cours des 2 ou 3 derniers tours permettrait de constater cette évolution en temps réel.

Les caractéristiques du jeu de Galèrapagos font également que les stratégies peuvent changer au cours du jeu en fonction de la situation, ce qui ajoute du bruit dans les données et des imprécisions dans la détection. De plus, les stratégies pragmatiques ne sont pas bien définies, étant souvent à l'interface entre les stratégies collaboratives et individualistes. Un autre jeu avec des stratégies plus stables et plus faciles à discriminer aurait donc été plus adapté pour une détection facile. Néanmoins, il est possible d'argumenter que dans ce sens, Galèrapagos reflète mieux les conditions réelles, dans lesquelles les frontières entre les différentes stratégies et postures adoptées sont souvent floues et évoluent.

Pour ce qui est des messages d'influence ou « protips », ils n'ont souvent pas retenu l'attention des joueurs, de par leur positionnement dans le jeu qui en fait un élément perturbateur pour certains. Une des pistes d'explication serait la tunnelisation de l'attention, qui pourrait introduire une résistance à la manipulation. Pour éviter ce syndrome de persévération, nous pourrions rechercher d'autres méthodes d'influence, par exemple enlever de l'information pour attirer l'attention de l'utilisateur, plutôt que d'en ajouter (Dehais et al., 2003).

Lorsque les participants ont lu les messages d'influence, ils ont pu les interpréter dans leur sens : par exemple, certains joueurs individualistes semblaient parfois interpréter une injonction à jouer de manière collaborative comme la nécessité de faire croire aux autres joueurs qu'ils sont collaboratifs, ceci afin de gagner leur confiance et pouvoir mieux les trahir par la suite. Nous remarquons que si l'influence semble avoir fonctionné au cours des parties dans certains cas, elle modifie rarement les tendances stratégiques profondes : il serait intéressant d'expérimenter d'autres méthodes d'influence et de les explorer sur le plus long terme.

Une autre limite de notre expérimentation est que le système Postdare est ici utilisé pour détecter la stratégie du joueur dans un contexte de problème fermé, en connaissant tous les paramètres à considérer et tous les choix du joueur, même ceux qu'il cache aux autres joueurs : par exemple, lorsqu'il conserve une carte individualiste sans la dévoiler aux autres ou lorsqu'il ment concernant ses cartes en main. Dans une véritable position d'adversité, il s'agirait d'un problème ouvert dont nous ne connaissons pas tous les paramètres. Nous disposerions sans doute de plus d'informations (collectées sur une plus longue durée), mais leur accès serait plus difficile. En effet, elles sont « cachées », non connues de l'adversaire, et représentent des indices forts qui ont souvent un poids important en tant que cues dans Postdare.

Cependant, nous estimons que la preuve de concept que nous avons mise au point a démontré son potentiel d'analyse rapide permettant de réagir en temps réel aux évolutions observées. Nous postulons

que son utilisation dans d'autres contextes pourrait apporter de nombreux avantages. Ainsi, un tel système pourrait être utilisé en cas de conflit pour détecter les actions de guerre cognitive ou encore monitorer les états de vulnérabilité des adversaires (voir *Chapitre 5*), et pour proposer des solutions de contre-mesures en fonction de divers signaux faibles et points de surveillance définis à l'avance. Il pourrait être employé pour la détection des vulnérabilités d'alliés afin de surveiller les informations auxquelles l'adversaire pourrait accéder et qu'il pourrait exploiter pour nuire, et proposer des solutions correctives. Nous pourrions également imaginer une utilisation individuelle pour monitorer ses propres performances de manière objective et les améliorer, dans les domaines des jeux, du sport, de la consommation énergétique, des stratégies d'investissement…

Notre expérimentation a permis de dégager différentes étapes méthodologiques nécessaires pour l'adaptation du système ANTICIPE. La première étape est de construire l'arbre de détection d'informations critiques, en listant les points de surveillance les plus importants et les différents signaux faibles ou cues qui permettent de les détecter. Cette étape nécessite de rassembler des experts du sujet traité et une recherche exhaustive sur les différentes sources de données qui peuvent être exploitées. L'arbre de décision obtenu doit ensuite être testé en conditions réelles pour valider les hypothèses, le corriger, l'enrichir et ajuster les poids associés aux signaux faibles. Étant la base du système, il doit être construit de manière très précise. Des simulations ou des wargames pourraient contribuer à améliorer l'arbre de détection d'informations critiques. Il est également nécessaire de rassembler des experts pour concevoir les différentes réponses possibles que le logiciel proposera : dans notre cas, elles étaient représentées par des messages d'influence, mais dans un contexte réel elles devront être élaborées et adaptées finement au contexte, avec une évaluation des résultats attendus.

Ainsi, nous étudierons dans le *Chapitre 5* les signaux faibles permettant d'identifier les individus les plus vulnérables et les plus importants qui peuvent être ciblés par des actions de guerre cognitive dans une organisation, posant une partie des bases pour construire l'arbre de détection d'informations critiques qui permettra d'utiliser le système ANTICIPE (Postdare) dans le cadre de la guerre cognitive.

PARTIE III – Étude prospective

Chapitre 5 - Proposition de modèle pour catégoriser les critères des cibles de guerre cognitive

5.1 Introduction

Après avoir étudié la guerre cognitive dans ses mécanismes, ses objectifs et ses effets sur la prise de décision, ce chapitre introduit une nouvelle dimension nécessaire à la conception d'un système de soutien à la décision dans un cadre de guerre cognitive : l'identification des cibles humaines. Nous donnons également des exemples d'outils et méthodes d'attaque, bien qu'ils ne soient pas exhaustifs – sur ce point, nous pourrons nous référer au paragraphe *1.6 Outils & armes de la guerre cognitive*.

La guerre cognitive peut agir sur chaque étape du processus décisionnel. Elle peut intervenir dès la formulation et la mise à jour d'une stratégie, en agissant sur la représentation des forces en présence, la créativité stratégique, la planification ou encore la mise en place du suivi informationnel adéquat à l'accomplissement de la stratégie établie et à sa modification si nécessaire. Elle peut également modifier la récolte et le traitement de l'information jugée utile pour la décision, la chaîne de communication de l'information et l'interprétation de l'information (sense-making). Enfin, elle peut viser directement la prise de décision à travers le rassemblement des informations nécessaires, l'évaluation des options possibles et de leurs conséquences, le choix d'une option ou encore la validation de la décision. Ces étapes sont autant de points de vulnérabilité susceptibles de faire l'objet d'actions de perturbation ou de manipulation qui peuvent être dirigées vers des individus clés.

Ainsi, dans une logique d'attaque comme de défense, il est important d'identifier les individus les plus exposés, les plus influençables ou encore les plus stratégiquement déterminants, que ce soit au sein de nos propres organisations ou chez l'adversaire. Dans le cadre de ce travail de thèse et de notre démarche de modélisation, il apparaît donc nécessaire de développer une typologie des cibles de la guerre cognitive. Ce chapitre vise à poser les fondations conceptuelles pour intégrer ce volet d'analyse dans un futur système de soutien à la prise de conscience de situation et à la prise de décision. Nous proposons alors de :

1) Qualifier les cibles vulnérables et valorisables chez l'adversaire aussi bien que les individus à protéger en priorité dans le camp allié ;
2) Recenser des signaux faibles potentiellement exploitables dans la construction d'un arbre de détection d'informations critiques. Celui-ci a vocation à être intégré au système technologique d'aide à la décision en contexte de guerre cognitive, dont cette thèse cherche à établir les fondations ;
3) Suggérer des exemples de stratégies à adopter pour certains profils de cibles identifiés.

Ce travail de recherche vise à répondre aux questions suivantes :

- Qu'est-ce qui fait qu'un individu est plus pertinent à cibler qu'un autre dans le cadre d'une stratégie de guerre cognitive ?

- Comment détecter les points de vulnérabilité dans la chaîne décisionnelle, et quelles méthodes permettent d'exploiter ces vulnérabilités ?

Il est important de noter que nous n'intégrons pas tous les critères et techniques existants dans le modèle présenté ici. Nous estimons que certaines sont peu recommandables dans un cadre démocratique. Cependant, nous les mentionnons car il est important de les connaître afin de s'en protéger.

Dans ce chapitre, nous utilisons le terme « organisation » pour désigner tout groupe humain structuré qui prend des décisions stratégiques et peut être visé par des attaques de guerre cognitive : Command and Control (C2), entreprise, état-major, gouvernement, association…

5.2 Méthode

Pour construire ce modèle, nous avons recueilli l'avis de spécialistes des opérations militaires et nous nous sommes inspirés d'expérimentations menées au sein des armées françaises et de l'OTAN (sources non communicables). En parallèle, nous avons mené une analyse de la littérature sur le sujet pour conforter le choix des critères retenus (présentés ci-après).

Nous nous sommes concentrés sur deux axes :

- Axe 1 : Les critères qui rendent un individu plus influent et donc plus intéressant à cibler au sein de l'organisation-cible ;
- Axe 2 : Les critères qui rendent un individu plus vulnérable aux attaques de guerre cognitive.

5.3 Modèles de ciblage documentés dans la littérature

Nous avons identifié dans la littérature différents modèles qui peuvent nous aider dans la construction de ce modèle de catégorisation des critères d'identification des cibles potentielles de guerre cognitive.

Target Audience Analysis (TAA), ou analyse du public-cible :

L'analyse du public cible est un outil utilisé par les producteurs de contenu, les professionnels des médias, les chercheurs et les organisations, visant à identifier et à comprendre les groupes spécifiques d'individus qu'ils souhaitent atteindre afin d'adapter leur production culturelle, leurs messages, leurs recherches ou leurs produits et assurer un effet maximal de ces derniers (Thompson & Weldon, 2022).

Ces techniques incluent (Thompson & Weldon, 2022 ; Camilleri, 2018 ; Williams, 2024 ; Smits, 2023) :

- L'analyse d'audience numérique utilise des systèmes qui collectent et analysent des données sur l'interaction du public avec du contenu sur diverses plateformes. Ces outils permettent de suivre les préférences et d'ajuster les stratégies en conséquence.
- La segmentation de l'audience consiste à diviser le public en groupes distincts selon des critères spécifiques : démographiques (âge, sexe, profession, statut marital, etc.), psychographiques (personnalité, valeurs, motivations, intérêts, mode de vie, etc.), géographique (localisation, climat, densité de population, etc.).
- Des mécanismes de rétroaction et de recherche qualitative permettent de mieux comprendre les besoins et préférences du public en recueillant des retours directs : entretiens, focus groups, sondages en ligne.

- Des techniques d'analyse avancée des données aident à comprendre les comportements, les préférences et prédire les tendances futures : analyse du comportement (interactions avec le contenu), analyse prédictive (anticipation des tendances futures à partir des données passées), analyse de sentiment (émotions et réponses affectives), analyse en temps réel (suivi des comportements via des technologies mobiles et la géolocalisation), *« digital trail »* (exploitation des données laissées par les utilisateurs en ligne et via des applications mobiles).
- L'approche analytique discursive examine comment un public construit socialement son comportement, en s'intéressant particulièrement à l'identité du public et à la manière dont il se perçoit et perçoit les autres : analyse des discours (significations produites par des discours liés à l'identité), analyse des différences ou similarités dans les discours et le langage utilisé, construction de l'identité…

Ces différentes techniques et les outils d'analyse utilisés peuvent nous inspirer pour identifier et qualifier les cibles potentielles de guerre cognitive.

MICE :

Le modèle MICE (Money, Ideology, Compromission / Coercition, Ego / Excitement) est un acronyme d'origine incertaine (KGB, FBI ?) désignant des facteurs de recrutement d'un espion (Petkus, 2010 ; Michalak, 2011) :

- Money : l'individu ciblé peut être intéressé par ou avoir besoin d'argent et ainsi être corrompu ou recruté pour une somme suffisamment élevée à ses yeux.
- Ideology : l'individu peut avoir une idéologie qui le rapproche de la cause de l'attaquant et rejoindre celle-ci par sa propre volonté ou parce qu'il y est appelé. Cet élément se rapproche des Valeurs dans le modèle décrit dans ce document.
- Compromission / Coercition : l'individu peut être forcé à rejoindre la cause par un Kompromat ou autre menace. Un individu peut par exemple être compromis en utilisant des personnes désirables comme appâts et des *« opérations Roméo »*, ou en exploitant ses comportements déviants (transactions illégales par exemple) ou perçus comme déviants par la société qui l'entoure (comme l'orientation sexuelle). Ces différents éléments peuvent engendrer chez lui de la honte ou la volonté de cacher l'information en question.
- Ego / Excitement : du côté de l'ego, l'individu peut être motivé par sa frustration due à un manque de reconnaissance, sa colère contre une organisation qui a piétiné son ego, ou encore être sensible à la flatterie. Du côté de l'*excitement*, se trouvent les personnalités de type T, qui sont caractérisées par leur recherche de nouveauté et d'excitation, qu'elle soit intellectuelle ou physique, la recherche de sensations fortes et peuvent éventuellement présenter des comportements déviants (toxicomanie et autres addictions).

La motivation à passer à l'acte est souvent causée par plusieurs facteurs, l'argent seul ne suffit généralement pas par exemple.

Ce modèle permet d'identifier des personnes vulnérables ou intéressées afin de recruter des agents. Il peut enrichir notre modèle mais n'a pas le même objectif : en effet, la guerre cognitive est généralement insidieuse et agit sans que la cible de l'attaque en soit consciente, ce qui rend une partie des critères proposés par le modèle MICE non valables ici (notamment Money et Compromission / Coercition).

5.4 Les principaux critères identifiés

Parmi les critères qui rendent un individu influent (Axe 1), nous avons déterminé qu'il était important de se pencher sur :

- le rôle de l'individu-cible dans le processus de Décision (D) ;
- la capacité d'Influence (I).

Parmi les critères qui rendent un individu plus vulnérable ou plus facile à cibler par des actions de guerre cognitive (Axe 2), nous proposons :

- la Volonté de coopération avec l'attaquant (V) ;
- l'Adaptation au changement (A) ;
- la Sensibilité à la désinformation et aux failles cognitives (S).

L'ensemble de ces critères constitue un modèle que nous avons nommé DIVAS (voir paragraphe *5.5 Modèle obtenu*). L'objectif de ce modèle est d'aider à identifier les points de vulnérabilité au sein d'une organisation afin de déstabiliser cette dernière ou influencer ses décisions dans une logique de guerre cognitive.

5.4.1 Décision : Rôle de l'individu dans le processus de prise de décision

L'accès et le rôle de l'individu dans la chaîne d'information et de décision (prise de l'information, communication de l'information, prise de décision et validation de la décision) est un critère primordial pour déterminer si l'influencer permettra effectivement de perturber l'organisation ciblée.

Le processus décisionnel peut être perturbé, bloqué ou modifié à toutes les étapes. Tout individu qui intervient dans la chaîne d'élaboration de la décision visée peut être pertinent à cibler, d'autant plus s'il est en capacité de valider ou d'influencer cette décision.

Selon Drafat (2023), si de nombreux acteurs peuvent être impliqués dans le processus, la prise de décision repose souvent sur un décideur principal clairement identifié et qui porte la responsabilité finale. Ainsi, ce décideur coordonne les rôles des autres acteurs et organise les flux d'information. Sa perception, ses valeurs et son pouvoir d'action influencent fortement la décision. Il peut aussi mobiliser la participation des acteurs comme levier stratégique, que ce soit pour améliorer la décision finale en recueillant leur avis, pour donner plus de légitimité à la décision, renforcer leur adhésion à la décision finale ou encore pour mieux diffuser l'information associée. Chaque participant au processus est en situation d'influencer activement la décision selon ses valeurs, expériences et expertises. Il participe à différentes étapes : définition du problème, évaluation des options, choix des critères, etc. Les collaborateurs sont impliqués de façon systémique dans le processus, en support aux décideurs, dans une logique de collaboration et de participation. L'entité collective (le groupe) ne doit pas être négligée : la décision est souvent le fruit d'une dynamique de groupe, surtout dans les grandes organisations. Les parties prenantes de la collectivité apportent des perspectives diverses, influencent le processus via des rapports de pouvoir, et leur implication peut varier en intensité et en durée. Enfin, les usagers ne participent pas directement à la décision mais en subissent les effets, leur satisfaction est donc une finalité recherchée.

Par ailleurs, Vroom et Yetton (1973) proposent une taxonomie des styles de décision dans les organisations selon la participation des subordonnés. Ils distinguent deux types de problèmes : individuels

(concernant un seul subordonné) et collectifs (concernant un groupe de subordonnés). Ils y appliquent alors quatre styles de décision possibles : autocratique (le leader prend la décision seul), consultatif (le leader consulte les subordonnés avant de prendre la décision), consensuel (le leader et les subordonnés prennent la décision en groupe) ou délégué (le leader délègue la prise de décision à un ou des subordonnés). Les auteurs établissent que, bien que le style de décision varie selon la situation, les managers tendent à recourir à des méthodes rapides, généralement parmi les options les moins participatives.

De façon complémentaire, le modèle juge-conseiller étudie la manière dont les décideurs intègrent les conseils donnés par leurs subordonnés. Cela dépend entre autres de la complexité de la tâche (qui augmente la prise en compte des conseils) et de l'expertise perçue du conseiller et la confiance en celui-ci (Pescetelli et al., 2021 ; Bonaccio & Dalal, 2006 ; Gino & Moore, 2007). La prise en compte des conseils n'est pas toujours optimale, le décideur ayant parfois tendance à sous-estimer les conseils externes par rapport à ses propres opinions (Yaniv & Kleinberger, 2000 ; Sniezek & Buckley, 1995).

Dans le milieu militaire, les décisions s'appuient sur des processus qui normalisent les interactions entre décideurs. Par exemple, hors temps de crise, l'armée de Terre française comprend trois assemblées qui mettent en application les décisions prises par le chef d'état-major[10] : le comité stratégique qui définit la ligne stratégique, le comité de commandement qui s'assure de la cohérence de l'action, et le comité exécutif qui s'occupe de la mise en œuvre des décisions. Comme dans les organisations civiles, le décideur porte la responsabilité de son choix (Charzat, 2024).

En temps de crise, le processus de décision militaire est différent pour incorporer une plus grande flexibilité et rapidité de décision. Il fonctionne sous forme de cycle continu de Command & Control (Figure 63) visant à :

- Planifier (**Plan**), c'est à dire élaborer ou itérer sur une stratégie ;
- Diriger l'action (**Direct**) afin d'assurer son déclenchement en temps voulu et sa réalisation de manière efficace ;
- Suivre son application (**Monitor**) ;
- Évaluer son exécution, sa pertinence et la stratégie planifiée à l'aide de son suivi et des résultats obtenus (**Assess**).

Ce processus repose sur une centralisation de la prise de décision stratégique, une coordination et communication permanentes, ainsi qu'une rétroaction et adaptation rapides. Il s'appuie sur un Working Group (groupe de travail), composé d'experts et de responsables, chargés de fournir des analyses, des recommandations et des solutions opérationnelles, et un Board de Décideurs, comité stratégique composé de hauts responsables militaires qui prennent les décisions majeures sur l'orientation générale de la réponse à la crise et assurent la cohérence de l'action avec les objectifs stratégiques (Claverie & Desclaux, 2022).

Ainsi, nous pouvons constater que, dans les domaines civils comme militaires, il existe des décisions prises plus ou moins collectivement et un unique décideur en porte généralement la responsabilité.

[10] https://www.defense.gouv.fr/operations

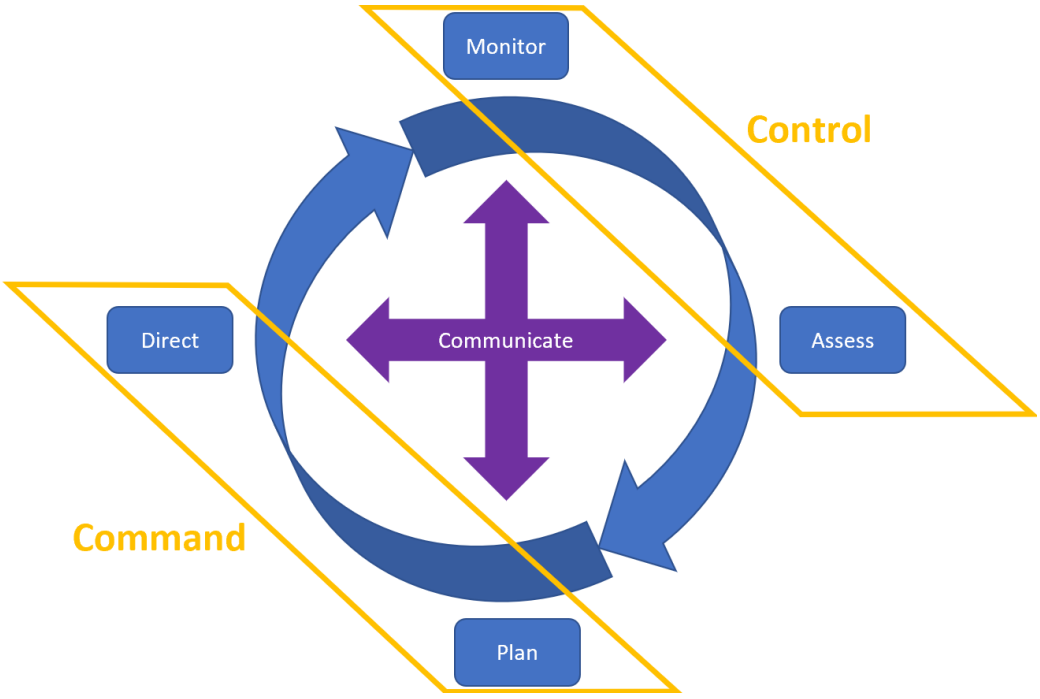

Figure 63 : Boucle du Command & Control militaire en situation de crise (inspirée de la doctrine militaire états-unienne du Command & Control – C2)

Sur la base de ces éléments, nous proposons dans le Tableau 27 une liste de statuts simplifiés qui permet de hiérarchiser par ordre d'importance croissante les personnes qui interviennent dans le processus décisionnel ou peuvent l'influencer. Ainsi, la personne qui « approuve » la décision et en porte la responsabilité a généralement la plus grande influence sur celle-ci, en particulier en temps de crise (Wen et al., 2025). Puis vient l'individu qui « décide » sous l'autorité du responsable. De nombreux acteurs interviennent dans le processus de décision au niveau de la prise et l'évaluation des informations : ils « estiment » l'information. Enfin viennent les usagers, qui « utilisent » l'information ou sur qui la décision aura un effet et dont l'existence a donc une influence sur cette décision sans qu'ils y prennent part ou en soient conscients.

Tableau 27 : Hiérarchisation des rôles dans la prise de Décision. Les individus généralement les moins pertinents à cibler sont en rouge, les plus pertinents en vert

D1	**Utilise :** L'individu peut faire partie du processus décisionnel mais de manière passive (partie prenante, usager…). Seront rangés dans ce critère les individus qui subissent les effets de la décision ou qui utilisent l'information sans donner leur avis la concernant, sans la modifier et sans la filtrer pour d'autres personnes, ainsi que les individus qui n'ont aucun accès à l'information. Il n'a pas d'impact sur la décision finale et n'est donc pas pertinent à cibler, sauf exceptions (activiste, nombre important d'individus qui réagissent à l'information…).
D2	**Estime :** L'individu est un acteur du processus de décision au niveau de la prise et l'évaluation des informations (analyste, expert, collaborateur…), il peut donc influencer la perception ou l'évaluation des décideurs en filtrant l'information ou en y associant un avis ou une expertise, perturbant la décision. Il peut donc être pertinent de le cibler pour orienter la perception de l'information et donc la décision.

	À long-terme, il pourrait aussi devenir un atout intéressant pour l'attaquant, notamment s'il est susceptible d'évoluer vers un poste plus important.
D3	**Décide :** L'individu fait partie du cercle de direction de l'organisation ou est en position de prise de décision sur certains sujets (intervenants / acteurs / experts). Il est un des acteurs de la décision visée et a donc une influence sur l'issue de celle-ci. Il est donc très intéressant à cibler et pourrait faire basculer les décisions en faveur de l'attaquant. Il pourrait devenir une carte maîtresse pour lui à plus long terme.
D4	**Approuve :** L'individu est à la tête de l'organisation (décideur) et il doit être visé en priorité. Il porte la responsabilité des décisions les plus importantes, il a donc droit de veto sur celles-ci et son maintien sous influence est primordial pour l'attaquant, pour garder le contrôle sur ses décisions.

Notre objectif principal est ici de distinguer les personnes qui ont une influence sur la décision et qui sont donc pertinentes à cibler pour un attaquant, de celles qui n'ont pas d'effet direct sur celle-ci et sont donc à prendre en compte mais moins importantes.

Nous soulignons l'importance de protéger les décideurs, ou lorsque c'est possible, de cibler ceux de l'adversaire. De nombreuses études ont montré que le décideur peut être influencé par divers facteurs tels que sa propre intégrité physique et physiologique, la maladie, la fatigue, des traumatismes etc. (Charzat, 2024 ; Gamble et al., 2018 ; Fooken & Schaffner, 2016).

Riggio et Newstead (2023) relèvent que les décideurs ne façonnent pas seulement les décisions, mais influencent directement la manière dont leurs équipes perçoivent, interprètent et traversent la crise. L'influence du décideur ne dépend pas uniquement de sa position hiérarchique, mais aussi de son aptitude à créer du sens, mobiliser les ressources cognitives du groupe et structurer l'action. Par ailleurs, les théories du pouvoir social et du leadership (French & Raven, 1959 ; Caillé, 2016) indiquent que l'individu en position de pouvoir exerce une influence sur son entourage sous trois formes. Premièrement, il est perçu comme ayant l'autorité pour punir ou récompenser, ce qui relève des pouvoirs de coercition et de récompense (leadership transactionnel). Deuxièmement, sa légitimité à diriger découle de sa position au sein de l'organisation et donc son pouvoir légitime (leadership institué traditionnel). Enfin, il sert de référence pour son entourage du fait de ses caractéristiques enviables. Ses pairs s'identifient à lui car ils appartiennent au même groupe de référence, d'autres voudraient s'attribuer ou acquérir ses caractéristiques, ce qui correspond aux pouvoirs expert et référentiel (leadership charismatique). Ainsi, la posture d'un décideur comme ses décisions ont une influence profonde sur le groupe qu'il dirige (Grint, 2005).

Il ne faut pas pour autant négliger les individus qui interviennent dans la décision, notamment en transmettant l'information (D2 dans le Tableau 27) : en effet, les décisions sont souvent le fruit d'une dynamique collective, et les interactions et les relations de pouvoir influencent la prise de décision (Castel & Chessel, 2024). De plus, la valeur de l'information dépend fortement de la confiance accordée à la source humaine qui la transmet (Rivière, 2017).

Il existe également des groupes d'influence qui exercent un rôle dans la décision de par leur existence, leur expertise et leur positionnement, sans en être les auteurs directs (Genieys et al., 2003). Ceux-ci peuvent également être des cibles clés pour un attaquant.

Quels sont les signaux faibles permettant de déterminer le statut de l'individu dans ce critère ?

Pour déterminer le statut de l'individu dans le critère de la Décision, il est possible de s'appuyer d'une part sur le renseignement humain, ou HUMINT, et d'autre part sur le renseignement technique, ou TECHINT (Hung & Hung, 2022). Le renseignement humain afin d'identifier sa position hiérarchique ou son poste dans l'organisation cible, par exemple via du renseignement en sources ouvertes (OSINT) ou en obtenant des informations de la part d'autres personnes dont l'accès est plus facile dans l'organisation, même si elles sont moins bien situées dans la chaîne de décision. Le renseignement technique peut également se montrer utile : écoute, localisation et déplacements qui permettent notamment de déduire qui l'individu rencontre, etc. (Williams et al., 2015).

Exemples de stratégies offensives basées sur ce critère :

Théories de la dissonance cognitive et prise de décision :

Selon la théorie fondatrice de la dissonance cognitive proposée par Leon Festinger (1957), toute décision crée une tension interne lorsque les alternatives écartées demeurent attractives, ce qui conduit l'individu à réévaluer ses choix afin de réduire cet inconfort. Ainsi, l'évaluation d'une décision est modifiée par la décision elle-même : une fois prise, elle est généralement perçue de manière plus positive, tandis que les options rejetées sont dévalorisées (Brehm, 1956 ; Harmon-Jones & Harmon-Jones, 2002). Il est à noter que plus les processus décisionnels ou négociations ayant abouti ont nécessité des concessions de la part de la personne, plus elle apprécie positivement ces accords ou décisions. Elle est donc susceptible de les reproduire par la suite (Aronson & Mills, 1959 ; Beauvois & Joule, 2020). Ce phénomène illustre le concept de justification de l'effort, paradigme évoqué dans l'introduction. Le fait d'être en position de responsabilité tend donc à ancrer des décisions comme étant fondamentalement bonnes et à s'exposer aux méthodes de manipulation et d'ingénierie sociale. En conséquence, il est possible d'analyser les décisions antérieures pour utiliser les concessions identifiées dans des négociations passées avec un individu cible pour renforcer la valeur de son résultat (de son point de vue).

Motivation à traiter l'information :

La position d'une personne dans la chaîne de commandement peut avoir un impact sur son implication dans la prise de décision pour diverses raisons (exemple : risque perçu de demande de justifications, pertinence pour sa tâche, responsabilité…). D'après l'Elaboration Likelyhood Model (ELM) développé par Petty et Briñol (2011), l'implication est un des facteurs influençant la probabilité qu'un individu étudie en détail l'argumentation d'une source d'information ou qu'il se base sur des indices périphériques, comme la crédibilité de la source ou l'alignement des idées avec ses propres opinions. De plus, Botelho et Coelho (1996) suggèrent que la motivation influence la profondeur du traitement de l'information, affectant ainsi la qualité des décisions. Par exemple, une motivation élevée peut conduire à une recherche d'information plus exhaustive et à une évaluation plus critique des options disponibles. Quant à la théorie de la focalisation réglementaire, elle distingue deux orientations motivationnelles : la recherche de gains et l'évitement des pertes (Higgins, 2011). L'attaquant peut utiliser ces éléments à son avantage pour influencer un analyste peu investi, qui pourrait alors mettre en avant un document parce qu'il corrobore ses opinions, ou amener un décideur à perdre du temps en lisant une argumentation complexe qui ne mène à rien. Une multitude de stratégies peuvent donc découler de cette information.

5.4.2 Influence : Capacité d'influence de l'individu

La possibilité d'influence d'un individu dans une organisation ne dépend pas uniquement de sa position hiérarchique formelle, mais aussi de sa place dans les réseaux relationnels (Lazega, 1994). Un acteur

central dans ces réseaux (par exemple dans des rôles de conseil, de contrôle informel ou de leadership) joue fréquemment un rôle déterminant dans la coordination, la circulation de l'information et la gouvernance. Cette centralité, qui structure l'interaction entre les membres de l'organisation, peut renforcer ou freiner les dynamiques collectives. Lazega (ibid.) distingue différents types de réseaux (d'entraide, de pouvoir, de conseil, etc.) qui sont parfois imbriqués mais pas nécessairement congruents. Ainsi, une personne peut être marginale dans un réseau de pouvoir tout en étant centrale dans un réseau informel d'entraide. Cela démontre la nécessité de dissocier l'influence réelle d'un individu de sa fonction officielle, sa position dans les réseaux pouvant révéler un accès privilégié à des ressources, à l'information, ou à des leviers d'action complémentaires à ceux des circuits formels de décision.

Cette perspective rejoint la notion de leadership informel, qui désigne l'influence exercée sans position de pouvoir désignée. Ce type de leadership émerge d'un processus social fondé sur la reconnaissance, le savoir et les interactions au sein du groupe. Leino (2022) montre que le leader informel peut stimuler la productivité, renforcer les dynamiques collectives, favoriser l'innovation et induire des changements organisationnels. Cette influence repose sur le savoir, l'action et la communication (Leino, 2022 ; Pescosolido, 2001). Un leader informel peut donc catalyser l'action collective, orienter les décisions et être une cible stratégique pour quiconque souhaite affaiblir une organisation ou y introduire du changement.

Nous proposons d'évaluer la capacité d'influence d'un individu à travers le nombre de contacts qu'il entretient au sein de l'organisation cible et de la chaîne de décision visée. Ce critère est composé de quatre statuts possibles : « sans contact », « quelques contacts », « nombreux contacts » et « influence profonde » (Tableau 28).

La grille de lecture proposée est volontairement souple, destinée à permettre une catégorisation affinée par l'analyse de la structure du réseau, de la centralité de l'individu et de la reconnaissance dont il bénéficie auprès de ses pairs. En effet, le nombre de contacts, comme la position hiérarchique, n'est pas un indicateur suffisant en soi. La théorie des trous structuraux (Burt, 2014) montre qu'un individu peu connecté peut jouer un rôle stratégique s'il relie deux sous-groupes isolés. En occupant cette position de pont, il contrôle le flux d'informations entre des parties du réseau ou des structures qui seraient autrement déconnectées, ce qui lui confère un pouvoir de médiation important. Il convient donc de ne pas négliger les individus qui n'ont pas nécessairement une position apparemment importante au sein de l'organisation ni une influence évidente sur la décision, à partir du moment où ces personnes peuvent être un point d'entrée et de contact pour l'agresseur.

Enfin, Freeman (1978) identifie trois dimensions fondamentales de la centralité dans un réseau : le degré (nombre de connexions), la proximité moyenne à tous les autres nœuds et la capacité à faire le lien entre des groupes distincts. Ces dimensions complémentaires permettent d'approcher plus finement le potentiel d'influence d'un individu, indépendamment de son statut formel.

Tableau 28 : Hiérarchisation des individus par leur influence au sein de l'organisation cible. Les individus généralement les moins pertinents à cibler sont en rouge, les plus pertinents en vert

I1	**Sans contact :** L'individu est isolé et n'a pas de contact avec le groupe visé, donc pas d'influence sur le groupe en question. Il n'est généralement pas pertinent pour l'attaquant de cibler cet individu. Même s'il a le potentiel de se rapprocher du groupe par la suite, il resterait peu pertinent de le prendre pour cible car cela demanderait des efforts considérables pour une faible espérance de gain.
I2	**Quelques contacts :** Avec seulement quelques contacts avec le groupe ciblé, l'individu a peu d'influence, il est peu utile. Il est peu pertinent de le cibler à moins qu'il ait le potentiel de se rapprocher du groupe ou de l'intégrer avec une forte probabilité, ou si aucun autre angle d'approche ne peut être identifié.
I3	**Nombreux contacts :** L'individu a un potentiel d'influence important. Avec de nombreux contacts au sein de l'organisation cible, il peut être utile pour l'attaquant de le viser. Nous pouvons également ranger dans ce critère les proches des personnes intervenant directement dans la chaîne de décision que l'attaquant cherche à perturber, étant donné que les liens familiaux et amicaux peuvent endormir les défenses cognitives et influencer efficacement les personnes visées (même sans le faire volontairement).
I4	**Influence profonde :** L'individu a de nombreux contacts au sein de l'organisation cible et a une influence profonde et bien ancrée sur les contacts en question. Il peut être par exemple un leader, un expert, un décideur, un influenceur ou toute autre personne vers qui les autres se tournent pour demander un avis en toute confiance. L'individu est alors très utile pour l'initiateur de l'attaque, à cibler en priorité.

Quels sont les signaux faibles permettant de déterminer le statut de l'individu dans ce critère ?

L'individu est-il dans l'organisation et a-t-il de nombreux contacts avec les individus visés ? Pour le déterminer, nous proposons de nous appuyer sur le renseignement humain et technique. Par exemple, il est possible de mener une analyse de texte (Natural Language Processing - NLP) sur ce que l'individu écrit ou ce qui se dit de lui sur les réseaux pour en savoir plus sur sa personnalité et sa réputation. En fonction des données disponibles, il pourrait être envisageable d'en déduire sa position hiérarchique et son rôle dans l'organisation, et savoir s'il est intégré à l'équipe visée (Rathi et al., 2022 ; Wen et al., 2025). Les réseaux alternatifs au sein de l'organisation peuvent également être étudiés, en identifiant si l'individu est impliqué dans des relations d'entraide, de bénévolat, avec des associations de travailleurs, etc. (Kojima, 2024). L'étude des interactions informelles, pendant une journée type comme lors d'événements de socialisation dans l'organisation, peut apporter des informations déterminantes sur sa capacité d'influence. Une autre piste et de tenter d'obtenir des informations de la part d'une source interne à l'organisation.

Si l'individu n'est pas membre de l'organisation, est-il un proche d'une personne visée et donc susceptible d'influencer indirectement la chaîne de décision ? Pour le savoir, l'analyste pourra étudier les sites de généalogie, les réseaux de contacts et d'interactions sur les réseaux sociaux, ou à nouveau tenter

d'obtenir des informations de la part d'une source interne à l'organisation (Zheng et al., 2024 ; Hristova et al., 2014).

Exemples de stratégies offensives basées sur ce critère :

Les personnes ayant peu d'influence apparente peuvent constituer un point d'entrée dans l'organisation ciblée. Toutefois, il est généralement plus utile de cibler directement les leaders formels ou informels, qui auront une influence profonde sur les membres de l'organisation, leur efficacité, leurs idées et leur aptitude ou volonté de changement.

5.4.3 Volonté : Volonté de coopération avec l'attaquant

Dans ce critère, qui peut être comparé au critère *« Ideology »* du modèle MICE (voir paragraphe *5.3 Modèles de ciblage documentés dans la littérature*), nous nous intéressons aux valeurs et convictions personnelles de l'individu et à sa posture envers l'agresseur. Est-il attaché et fidèle à son organisation ? Y a-t-il une dissonance entre ses valeurs et celles de son organisation ? Est-il prêt à s'intéresser à la cause de l'attaquant ?

La théorie des comportements opportunistes prédit que les acteurs cherchent à maximiser leur autonomie en exploitant les zones d'incertitude (Crozier & Friedberg, 1977). Un individu peut faire preuve d'opportunisme actif (mensonge, tromperie) ou passif (rétention d'information) pour maximiser ses gains individuels, même au détriment de la coopération (Fulconis & Paché, 2008). Ainsi, un individu peut adopter un comportement de saboteur ou de supporteur au sein de l'organisation en fonction de ses intérêts et des opportunités perçues.

Par ailleurs, Amblard et al. (2005) introduisent la notion des logiques d'action, reconnaissant que les comportements peuvent être guidés par diverses rationalités (stratégique, identitaire, culturelle, etc.). Le contexte (opportunités offertes par l'organisation, interactions avec les autres acteurs…) et les constructions identitaires permettent donc de comprendre comment les individus peuvent adopter des comportements variés, du sabotage au soutien en passant par l'opportunisme.

Le concept de la fenêtre d'Overton (Russell, 2006) montre qu'un comportement ou une idée considéré(e) comme acceptable par l'opinion publique à un moment donné peut devenir extrême ou inacceptable par un glissement des valeurs, et inversement. Cette théorie peut être transposée au sein d'une organisation, où certaines valeurs peuvent devenir acceptables peu à peu et ainsi rapprocher l'ensemble des individus concernés de la cause de l'attaquant.

Sur cette base de réflexion, nous proposons (Tableau 29) une hiérarchisation de « saboteur » (individu activement opposé à l'attaquant) à « supporteur » (individu réagissant en faveur de l'attaquant en cas de sollicitation), en passant par l'individu « neutre » et « l'opportuniste », qui réagit contre l'attaquant en cas de sollicitation.

Tableau 29 : Hiérarchisation des individus par leur volonté de coopération avec l'initiateur de l'attaque. Les individus généralement les plus difficiles à cibler sont en rouge, les plus faciles en vert

V1	**Saboteur :** L'individu est activement opposé à l'attaquant. Pour l'attaquant, le rallier à sa cause demanderait de nombreux efforts et un temps considérable, sans réussite garantie car cela pourrait induire un effet de réactance. Il s'agit d'un individu à éviter de cibler, sauf dans des stratégies de polarisation de l'opinion où l'attaquant pourrait vouloir l'amener à des comportements encore plus extrêmes afin de *« ridiculiser »* sa propre cause.
V2	**Opportuniste :** L'individu réagit contre l'attaquant en cas de sollicitation. Il n'est pas directement une menace tant qu'il n'est pas provoqué. Pour l'attaquant, il vaut mieux éviter de le cibler trop frontalement pour éviter tout effet de réactance. Cependant, il est possible de mettre en place des actions subtiles ou le faire douter de sa cause pour qu'il soit moins susceptible de devenir actif contre l'attaquant.
V3	**Neutre :** L'individu reste neutre. Il est intéressant voire important pour l'attaquant de tenter de le rallier à sa cause avant qu'il se range à l'opinion opposée.
V4	**Supporteur :** L'individu réagit en faveur de l'attaquant en cas de sollicitation. Il est alors une arme d'influence de choix pour pénétrer l'organisation cible.

Quels sont les signaux faibles permettant de déterminer le statut de l'individu dans ce critère ?

Pour déterminer le statut de l'individu dans le critère de la Volonté de coopération avec l'attaquant, une piste importante est d'évaluer son positionnement sur les réseaux sociaux (messages postés, positionnement des personnes suivies, mots-clés identifiés) (Hristova et al., 2014). Il est également intéressant d'obtenir des informations de la part de sources proches de l'individu dans sa sphère privée.

Exemples de stratégies offensives basées sur ce critère :

Il existe différentes méthodes pour changer la perception d'un individu au sein d'une organisation, gains ou risques potentiels. Outre la corruption, la coercition ou autres méthodes peu recommandables, un autre levier disponible s'appuie sur les valeurs. En effet, si les valeurs de la personne cible sont très différentes de celles de son organisation (ou même lui semblent l'être), l'attaquant peut utiliser cette dissonance pour faire levier et le convaincre de coopérer (Cherré et al., 2014 ; Harmon-Jones & Harmon-Jones, 2007). Le kompromat (voir paragraphe *5.6.3 Stratégies d'attaque en fonction du type de cible*) peut également s'avérer efficace si des révélations faites sur ses valeurs ou les actions qui en découlent s'éloignent fortement de la fenêtre d'Overton de son organisation (Russell, 2006 ; Hübert & Little, 2020).

Dans tous les cas, si l'individu est déjà favorable à la cause de l'attaquant, il sera plus facile d'obtenir une coopération de sa part. Il sera possible de l'atteindre puis l'amener à une coopération progressivement de plus en plus importante.

5.4.4 Adaptation : Disposition de l'individu à s'adapter au changement

L'individu est-il intéressé par un changement de point de vue, prêt à s'adapter, ou au contraire souhaite-t-il conserver l'ordre social, politique ou militaire en place ?

La résistance au changement et sa réduction au sein des organisations et dans différents domaines comme la santé ou l'éducation a été largement étudiée (Metz, 2021 ; González et al., 2022). Oreg (2003) identifie quatre facteurs clés de résistance au changement : la recherche de routine, la réaction émotionnelle au changement imposé, la rigidité cognitive et la focalisation à court terme. Agocs (1997) établit que la résistance au changement dans les organisations peut se manifester sous différentes formes plus ou moins actives : déni de la légitimité du changement, refus de reconnaître sa responsabilité à changer, refus d'implémenter le changement adopté par l'organisation, sabotage du changement une fois que l'implémentation a démarré… En résumé : déni, inaction, répression. Oliver (1991) propose cinq types de stratégies de réponse aux pressions institutionnelles : consentement, compromis, évitement, défi et manipulation. À noter que le changement individuel suit également un processus en plusieurs étapes : choc, résistance, exploration et implication (Bridges, 1991).

Nous observons une grande variété de réactions au changement proposé ou imposé. En nous inspirant des travaux sur l'adaptation à l'innovation (Rogers, 1995 ; Orr, 2003), nous proposons une catégorisation des individus en fonction de leur disposition au changement (Tableau 30) : les « innovateurs » peuvent être rapprochés de ceux de la courbe de diffusion de l'innovation, qui sont *« aventuriers »* et les premiers à adopter cette dernière. Les « pragmatiques » seraient les adopteurs précoces et la majorité précoce, qui suivent les innovateurs une fois qu'ils ont adopté l'innovation. Les « conservateurs » et les « retardataires » correspondent respectivement à la majorité tardive et aux derniers adopteurs, qui sont les plus méfiants envers la nouveauté.

Tableau 30 : Hiérarchisation des individus par leur disposition à s'adapter au changement. Les individus généralement les plus difficiles à cibler sont en rouge, les plus faciles en vert

A1	**Retardataire :** L'individu est totalement réfractaire au changement. Il se laissera donc difficilement influencer et de changer d'opinion, il est alors peu pertinent à cibler. Cependant, il pourrait être manipulé pour mener inconsciemment des actions qui servent la cause de l'attaquant.
A2	**Conservateur :** L'individu agit pour maintenir l'ordre social et politique établi. Comme le Retardataire, il sera difficile de le faire changer d'opinion mais il peut tout de même être influencé pour mener des actions choisies.
A3	**Pragmatique :** L'individu est prêt à revoir et adapter ses valeurs en fonction de la situation. Il choisira probablement le camp qui lui est le plus favorable à court ou à long-terme. Il s'agit d'une personne qui peut être convaincue avec des arguments appropriés.
A4	**Innovateur :** L'individu est enthousiaste concernant le changement, il est parmi les premiers à l'adopter voire un initiateur du changement. Il est ouvert d'esprit et s'intéressera probablement à la cause de l'attaquant si elle lui est présentée de manière suffisamment convaincante. Il peut alors devenir un allié de choix pour l'agresseur.

Si l'individu est prêt à changer ou au moins s'intéresse à d'autres points de vue, il peut être considéré comme plus susceptible d'être confronté à des tentatives d'influence et pourrait même être consciemment rallié à la cause de l'attaquant si elle correspond à ses valeurs. À l'inverse, il pourrait aussi être plus facilement conscient de tentatives de manipulation.

Quels sont les signaux faibles permettant de déterminer le statut de l'individu dans ce critère ?

L'âge, le profil socio-économique, l'ancienneté dans l'organisation, l'engagement envers l'organisation et les relations de l'individu peuvent donner un niveau de probabilité de la disposition au changement, même s'il n'est pas très fiable. Les éléments de son histoire personnelle peuvent aussi fournir des indications, notamment les changements dans sa vie personnelle et professionnelle, ses activités etc. (Madsen et al., 2006).

Le positionnement de l'individu sur les réseaux sociaux et les personnes qu'il suit peut être étudié, en menant une analyse de texte (NLP) sur ce que l'individu écrit afin d'en savoir plus sur ses émotions, intentions et comportements, susceptibles de fournir des indices sur sa disposition à s'adapter au changement (Voelker et al., 2011).

Par ailleurs, une évaluation par contact direct avec la personne, ou avec une source proche de l'individu dans sa sphère privée ou professionnelle, peut être particulièrement informative.

L'échelle de la résistance au changement (Oreg, 2003) ou d'autres outils de mesure peuvent être utilisés lorsqu'il est envisageable de faire remplir des questionnaires à l'individu évalué.

Exemples de stratégies offensives basées sur ce critère :

L'exemple de la courbe de diffusion de l'innovation (Rogers, 1995) montre qu'à partir du moment où les innovateurs et les adopteurs précoces ou pragmatiques sont convaincus, ils ont tendance à entraîner le changement de la majorité. Il est donc stratégique de cibler ces individus pour générer une évolution des mentalités et positionnements au sein d'une organisation.

5.4.5 Sensibilité : Sensibilité à la désinformation et aux failles cognitives de l'individu

L'individu est-il sensible à la manipulation ? Est-il facile à manipuler au regard de ses failles cognitives ?

Piksa et al. (2022) identifient quatre *« phénotypes »* distincts de susceptibilité à la désinformation et établit leurs profils cognitifs et psychologiques : les plus fréquents sont les *« consommateurs »* ou *« crédules »* (*consumers*) et les *« douteurs »* ou *« sceptiques »* (*doubters*), qui ont tendance respectivement à juger toute information vraie ou toute information fausse indépendamment de sa véracité. Les deux autres profils sont moins fréquents : les *« incompétents »* (*duffers*) ont un faible discernement et les *« savants »* (*knowers*) évaluent bien la véracité de l'information.

Différents facteurs peuvent influencer la sensibilité à la désinformation. Par exemple, l'Ofcom a publié en 2025 une revue de littérature identifiant plusieurs facteurs tels que (Ofcom, 2025) :

- la littératie numérique et médiatique (une meilleure maîtrise des outils numériques et des médias est associée à une moindre susceptibilité à la désinformation) ;
- la confiance dans les institutions (une faible confiance dans les institutions gouvernementales et scientifiques est corrélée à une plus grande vulnérabilité à la désinformation) ;

- le niveau d'éducation (un niveau d'éducation plus élevé est généralement lié à une meilleure capacité à discerner les informations fiables ; de même, les personnes *« ayant un statut social bas en raison de leur ethnicité, revenus ou autres facteurs connexes sont plus susceptibles d'adhérer aux théories du complot »*) ;
- la culture, l'idéologie politique ;
- les traits de personnalité ;
- les connaissances préalables sur le sujet…

Le critère *« Ego / Excitement »* du modèle MICE (voir paragraphe *5.3 Modèles de ciblage documentés dans la littérature*) pourrait également constituer un type de faille cognitive : frustration, colère, sensibilité à la flatterie (Petkus, 2010 ; Michalak, 2011) peuvent être considérer comme des failles cognitives ou des traits de personnalité qui peuvent être exploités par un adversaire.

Tableau 31 : Hiérarchisation des individus par leur sensibilité à la désinformation et aux failles cognitives. Les individus généralement les plus difficiles à cibler sont en rouge, les plus faciles en vert

S1	**Entraîné :** L'individu est formé et informé sur les biais de la manipulation de l'information. Il suit des processus clairs et prédéfinis pour éviter les biais. Ce type d'individu est peu vulnérable à l'influence et la manipulation et n'est donc pas le plus pertinent à cibler.
S2	**Conscient :** L'individu est conscient de l'existence des failles cognitives ou méfiant envers l'information non crédule face à la désinformation. Ce type d'individu est plus difficile à exploiter mais reste une cible potentielle pour des attaques suffisamment subtiles et difficiles à détecter.
S3	**Crédule :** L'individu n'est pas conscient de ses failles cognitives potentielles et des campagnes de désinformation dont il peut faire l'objet. Ce type d'individu est plus facile à cibler avec des attaques de guerre cognitive, mais avec de l'aide ou de la formation il peut aussi passer dans le critère « conscient ».
S4	**Complotiste :** L'individu est sensible aux influences et aux biais cognitifs, il a peu d'esprit critique. Cela peut se traduire par une adhésion rapide à des informations non vérifiées ou par une méfiance déraisonnable vis-à-vis des institutions ou de son organisation, pouvant induire du complotisme. L'individu est facile à manipuler et peut être mené dans un déni de la réalité. Il pourrait ainsi être utilisé comme *« idiot utile »* et servir les intérêts de l'attaquant, consciemment ou non, même au détriment de ses propres intérêts (Claverie, 2023).

Ici, nous nous intéressons en particulier à ce qui rend un individu plus sensible à la désinformation et à la manipulation, et donc plus susceptible d'être visé par une attaque. Nous proposons donc une hiérarchisation en 4 classes des individus dans ce critère « sensibilité » (Tableau 31) : les individus les moins sensibles à la désinformation seraient les *« savants »* dans l'échelle de Piksa et al. (2022), que nous appelons ici « entraînés » : ils sont conscients des biais, leur esprit critique est entraîné (soit par leur niveau d'éducation et environnement social, soit par une formation intentionnelle). Viennent ensuite les « conscients », qui connaissent l'existence de leurs biais cognitifs et de la désinformation mais n'ont pas d'entraînement spécifique ; cette classe contient également les *« sceptiques »*, leur méfiance les rendant moins perméables aux stratégies d'influence. La troisième classe, « crédule », comprend les individus plus

crédules et faciles à désinformer. Enfin, la dernière classe, « complotiste », désigne les individus les plus susceptibles de croire en des théories du complot : ceux qui ont un faible niveau d'éducation et social et ont peu confiance envers les institutions.

List et al. (2024) ont montré que donner à des personnes des informations sur leurs propres biais n'avait pas d'impact sur leur scepticisme, mais leur montrer des vidéos de débiaisage améliorait leur esprit critique et les rendait plus prudentes envers la désinformation potentielle. Cela justifie l'intérêt de la formation et de l'entraînement à la détection de tentatives de manipulation.

Quels sont les signaux faibles permettant de déterminer le statut de l'individu dans ce critère ?

Pour déterminer si un individu est sensible à la désinformation, il est possible de prendre exemple sur le questionnaire MIST développé par l'Université de Cambridge (Maertens et al., 2023). Nous pouvons arguer de sa viabilité dans le temps étant donné qu'il est très dépendant de l'actualité et de l'opinion publique. Néanmoins, il constitue à la fois une preuve de concept et un exemple de questionnaire qui montre qu'il est possible de détecter la sensibilité à la désinformation : c'est un modèle sur lequel nous pouvons en construire de nouveaux. Un questionnaire ne peut généralement être donné qu'à ses alliés, à moins de réussir à piéger la personne-cible. Il reste un outil intéressant pour évaluer nos propres vulnérabilités. Piksa et al. (2022) ont utilisé un outil similaire pour évaluer la sensibilité à la désinformation, reprenant 24 titres d'actualités liées à la pandémie de COVID-19.

Les diverses fragilités et failles cognitives de l'individu se traduisent par son niveau de vulnérabilité psychologique ou émotionnelle, sa résilience au stress, sa mémoire et son attention qui permettent de mieux détecter les incohérences, sa flexibilité ou rigidité cognitive, son empathie (Vaillancourt, 2021 ; Rademacher et al., 2023).

Un niveau d'éducation élevé peut laisser supposer une moins grande perméabilité à la désinformation et aux failles cognitives. La littératie numérique et médiatique ainsi que le niveau de formation et d'information sur ces sujets sont de meilleurs indicateurs (Ofcom, 2025). À l'inverse, une personne ayant un faible niveau d'éducation et de connaissances sur le sujet traité est souvent plus facile à manipuler. Un autre facteur à évaluer est l'environnement social (famille, amis).

Un autre indicateur est la présence de comportements à risques, de comportements déviants ou à la limite de la légalité, mais aussi de comportements à risques en ligne, notamment les individus très actifs sur les réseaux sociaux, sur des plateformes de discussion ou sur des forums (Anaut, 2015 ; Simion & Dorard, 2020).

Exemples de stratégies offensives basées sur ce critère :

Failles cognitives :

De nombreuses failles cognitives sont exploitables dans un contexte de guerre cognitive ou d'influence, telles que les biais du raisonnement et de la décision, qui ont été présentés dans le paragraphe *2.2.2 Les biais de la décision*. Ces biais peuvent être exploités par un acteur malveillant, qui influencerait une situation ou une manière de présenter les informations afin de les rendre propices à la manifestation de ces biais (Dyèvre, 2015). Un exemple est l'exploitation des algorithmes de réseaux sociaux par un acteur qui en comprend les rouages, ou leur manipulation par un acteur qui les contrôle, afin d'influencer les perceptions ou les comportements des utilisateurs (Margraff, 2024).

Idiot utile :

Le terme d'idiot utile est attribué à Lénine, désignant initialement les intellectuels occidentaux qui ont été séduits par le régime soviétique et en ont fait la propagande. Ce concept désigne une personne naïve qui fait inconsciemment des actions favorables à l'ennemi (actions, propagande) aux dépens de ses propres intérêts (Facal, 2011 ; Greenslade, 2020 ; Claverie, 2023). Ce type d'action peut être mené envers les personnes les plus sensibles à la désinformation ou opposées aux institutions, qui peuvent alors servir de levier pour diffuser des idées ou obtenir un avantage stratégique.

Autres techniques :

D'autres pistes d'influence exploitent certains mécanismes cognitifs : par exemple, l'activation d'automatismes, le déclenchement de réflexes émotionnels par évocation de souvenirs, ou encore la manipulation des processus d'induction et d'abduction afin d'orienter les inférences produites. Ces procédés peuvent être renforcés par la surcharge cognitive ou informationnelle et la distraction, qui réduisent les capacités de contrôle et de vérification critique (Claverie & du Cluzel, 2021 ; Isaac et al., 2007).

5.5 Modèle obtenu

Comme présenté dans les paragraphes précédents, nous avons pu rassembler cinq critères comme importants à détecter dans l'organisation ciblée :

- le rôle dans le processus de Décision (D) ;
- la capacité d'Influence (I) ;
- la Volonté de coopération avec l'attaquant (V) ;
- l'Adaptation au changement (A) ;
- la Sensibilité à la désinformation et aux failles cognitives (S).

Pour chacun des critères, nous avons défini quatre statuts possibles de l'individu étudié, du moins pertinent ou plus difficile à cibler (en rouge dans le Tableau 32) au plus pertinent ou plus facile à cibler (en vert).

Ainsi, pour le critère « rôle dans le processus de prise de décision » (D), les statuts possibles classés du moins influent au plus influent sont : Utilise (utilise l'information), Estime (estime l'information et peut potentiellement la bloquer), Décide (est en position de décision) et Approuve (est en position de valider la décision).

Pour le critère « capacité d'influence » (I), les statuts proposés sont, du moins influent au plus influent : Sans contact (isolé, n'a pas d'influence sur le groupe), Quelques contacts (a peu d'influence, peu utile), Nombreux contacts (a un potentiel d'influence important) et Influence profonde (potentiel d'influence très important, modèle ou figure d'autorité de nombreux autres).

Pour le critère « volonté de coopération avec l'initiateur de l'attaque » (V), les statuts identifiés sont, du plus difficile au plus facile à cibler : Saboteur (activement opposé à l'initiateur de l'attaque), Opportuniste (réagit contre l'initiateur de l'attaque en cas de sollicitation), Neutre (reste neutre face à l'initiateur de l'attaque) et Supporteur (réagit en faveur de l'initiateur de l'attaque en cas de sollicitation).

Pour le critère « disposition à s'adapter au changement » (A), nous proposons les statuts suivants, du plus difficile au plus facile à cibler : Réfractaire au changement (non susceptible de changer), Conservateur (agit pour maintenir l'ordre social et politique établi), Pragmatique (prêt à revoir et adapter ses valeurs en fonction de la situation) et Innovateur (enthousiaste concernant le changement, parmi les premiers à l'adopter).

Tableau 32 : Modèle DIVAS : qualification des cibles potentielles de guerre cognitive. Les individus les plus pertinents ou faciles à cibler selon chaque critère sont en vert, les moins utiles ou plus difficiles à cibler en rouge

Critère		Statut	Définition du statut	Code
D	Rôle dans le processus de prise de décision	Utilise	Utilise l'information	D1
		Estime	Estime l'information et peut potentiellement la bloquer	D2
		Décide	En position de décision	D3
		Approuve	En position de validation de la décision	D4
I	Capacité d'influence	Sans contact	Isolé, pas d'influence sur le groupe	I1
		Quelques contacts	Peu d'influence, peu utile	I2
		Nombreux contacts	Potentiel d'influence important	I3
		Influence profonde	Potentiel d'influence très important, modèle ou figure d'autorité de nombreux autres	I4
V	Volonté de coopération avec l'attaquant	Saboteur	Activement opposé à l'initiateur de l'attaque	V1
		Opportuniste	Réagit contre l'initiateur de l'attaque en cas de sollicitation	V2
		Neutre	Reste neutre face à l'initiateur de l'attaque	V3
		Supporteur	Réagit en faveur de l'initiateur de l'attaque en cas de sollicitation	V4
A	Disposition à s'adapter au changement	Réfractaire au changement	Non susceptible de changer	A1
		Conservateur	Agit pour maintenir l'ordre social et politique établi	A2
		Pragmatique	Prêt à revoir et adapter ses valeurs en fonction de la situation	A3
		Innovateur	Enthousiaste concernant le changement, parmi les premiers à l'adopter	A4
S	Sensibilité à la désinformation et aux failles cognitives	Entraîné	Formé/informé sur les biais et la manipulation de l'information. Suit des processus pour éviter les biais	S1
		Conscient	Conscient de l'existence des biais ou sceptique face à l'information	S2
		Crédule	Non conscient de ses biais potentiels et des manipulations de l'information	S3
		Complotiste	Manque d'esprit critique, méfiance déraisonnable voire défiance envers les institutions ou son organisation	S4

Enfin, pour le critère « sensibilité à la désinformation et aux failles cognitives » (S), nous distinguons les statuts suivants, du plus difficile au plus facile à cibler : Entraîné (formé/informé sur les biais et la manipulation de l'information, suit des processus pour éviter les biais), Conscient (conscient de l'existence des biais et ou sceptique face à l'information), Crédule (non conscient de ses biais potentiels et des manipulations de l'information) et Complotiste (tendance au complotisme, déraisonnablement méfiant voire défiant envers les institutions ou son organisation).

Il faudra identifier quel est le statut des individus ciblés pour chacun de ces cinq critères afin d'en déduire quelles personnes cibles il serait le plus pertinent et utile de tenter d'influencer pour un attaquant.

5.6 Méthodologie de ciblage de l'adversaire

5.6.1 Utilisation du modèle DIVAS

Le modèle DIVAS peut être utilisé pour représenter schématiquement une organisation-cible, et ainsi mieux repérer les points de vulnérabilité de cette organisation, qui seront à renforcer ou à attaquer en priorité en fonction de l'objectif de la personne qui l'utilise (défense ou attaque). La structure des relations au sein de l'organisation aussi bien que la qualification de chaque personnage via les différents critères (D, I, V, A, S) peuvent donner des indices sur les stratégies qui peuvent être menées.

La Figure 64 donne un exemple de qualification de certains individus au sein d'une organisation fictive. Les caractéristiques démographiques des personnages représentés (prénom, sexe, ethnie) ont été générés aléatoirement et leur position dans l'organisation et leurs statuts selon chaque critère ont été choisis de manière arbitraire.

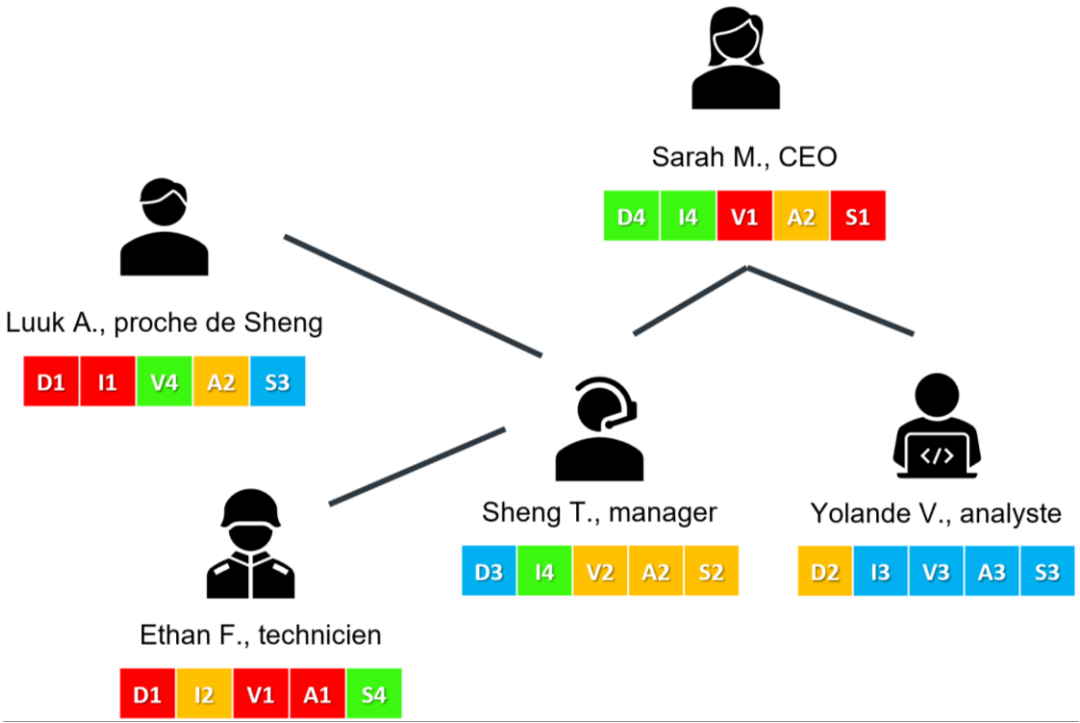

Figure 64 : Exemple de qualification des cibles potentielles avec le modèle DIVAS dans une organisation fictive

NB : Le schéma et les personnages présentés en Figure 64 ne reflètent aucune organisation réelle et toute ressemblance serait fortuite.

Comment cibler les individus décrits par le modèle ?

Nous donnons ci-dessous des exemples reprenant l'organisation fictive décrite dans la Figure 64.

1) Sarah M., CEO, est à la tête de l'organisation, ce qui lui donne une influence profonde sur les décisions et sur les collaborateurs (D4, I4). Cependant, elle est en concurrence avec l'initiateur de l'attaque (V1). Elle n'est pas ouverte au changement (A2) et a des défenses cognitives importantes (S1). Elle participe régulièrement à des événements autour de la désinformation et de la formation à l'esprit critique. Ces caractéristiques la rendent difficile à prendre pour cible de manière subtile à moins d'agir sur le long terme. Il sera plus facile de perturber d'autres maillons de la chaîne d'information et de décision et ainsi tenter de la déstabiliser de manière détournée, par exemple en la surchargeant d'informations inutiles qui la fatigueront et peuvent ainsi bloquer un processus décisionnel ou lui faire perdre ses repères en termes de valeurs.

2) Sheng T., manager, est un décideur influent (D3) sur certains sujets concernant les équipes et les projets qu'il gère (dont un qui intéresse en particulier l'attaquant). En plus du contrôle direct qu'il a sur son équipe, il a été identifié comme leader informel dont les compétences sociales sont reconnues et appréciées, lui conférant une influence importante au-delà de son poste (I4). Il a tendance à réagir négativement face à l'attaquant lorsqu'il est sollicité sur le sujet, mais ne lui est pas activement opposé (V2). Il a tendance à être conservateur et fidèle à son entreprise (A2), et très méfiant envers toute information qui lui est soumise, vraie ou fausse (S2). Sheng est plus facile à atteindre directement que Sarah mais reste une cible difficile. Malgré tout, son influence à la fois sur les décisions et sur les membres de l'organisation en font une cible très intéressante.

3) Yolande V., analyste, est, du fait de son métier, en position d'évaluer l'information (D2). Elle est avenante et a de nombreux contacts au sein de l'organisation (I3). Sur les réseaux sociaux comme au travail, elle mentionne peu ou pas l'initiateur de l'attaque et n'exprime pas d'opinion. Nous en déduisons qu'elle est neutre envers lui (V3). Elle a un caractère ouvert et est prête à changer (pragmatique), en témoignent ses changements d'employeur et de localisation relativement fréquents (A3). D'après son comportement sur les réseaux sociaux (posts *« likés »* et *« repostés »*), elle semble crédule et peu consciente de la désinformation et de ses biais (S3). Elle pourra alors être facilement ciblée par des méthodes de manipulation de l'information, notamment via les réseaux sociaux et éventuellement de l'ingénierie sociale, pour l'influencer à devenir plus favorable à la cause de l'attaquant. Ainsi, suivant le type d'informations auxquelles elle a accès, il est possible qu'inconsciemment, elle labellise, bloque ou transmette certaines informations d'une manière qui soit favorable à l'attaquant.

4) Ethan F., technicien, n'intervient pas dans les processus de décision ou dans l'évaluation des informations (D1). Il est réfractaire au changement (A1) et très actif contre l'initiateur de l'attaque (V1). Il a quelques contacts au sein de l'organisation, interagissant surtout avec d'autres membres de son équipe et du manager (I2). Il est particulièrement vulnérable à la désinformation et aux failles cognitives comparé aux autres collaborateurs (S4). Il peut alors être pris pour cible pour rendre ses positions encore plus opposées à celles de l'attaquant afin de cliver les opinions au sein de l'organisation, et éventuellement faire passer pour ridicule son comportement réfractaire au changement qui peut alors être vu comme extrémiste. L'attaque peut être accompagnée d'une stratégie pour influencer les opinions d'autres personnes dans l'organisation dans un sens qui est plus favorable à l'attaquant, afin de rendre le clivage plus efficace et détériorer l'ambiance de travail. Cela peut avoir des conséquences négatives sur l'efficacité de l'équipe voire de l'organisation et sur la qualité des décisions prises. Une telle attaque pourrait avoir un effet d'autant plus important s'il existe un grand nombre de « Ethan » dans l'organisation considérée.

5) Luuk A., proche du manager Sheng, le voit presque tous les jours dans un cadre privé, comme nous pouvons le voir sur les photos qu'il partage sur les réseaux sociaux. Étant en-dehors de l'organisation, il ne connaît que peu d'informations et dans tous les cas n'intervient pas sur celles-ci ni sur les décisions (D1). Son seul contact avec l'organisation est via Sheng, nous pouvons donc considérer qu'il est sans contact (I1). Il est crédule, non conscient de ses biais potentiels (S3). Il est relativement conservateur dans sa vision du changement (A2) mais est déjà favorable à l'attaquant, étant ouvertement un supporteur de sa cause (V4). Il pourrait donc être apparenté à un *« idiot utile »* et constituer un point d'entrée pour tenter d'affecter le manager et peut-être introduire des doutes chez lui en agissant dans la durée, dans un cadre privé où il pourrait baisser ses gardes ou se compromettre.

Comment cibler la décision dans l'organisation décrite par le modèle ?

L'objectif de l'attaquant est d'influencer la décision au sein de l'organisation fictive décrite précédemment (Figure 64).

Suite à l'analyse des différentes cibles potentielles, il apparaît que la stratégie la plus efficace (en rouge sur la Figure 65) consiste à cibler Yolande, analyste au sein de l'organisation. En effet, son profil la rend particulièrement vulnérable aux techniques de manipulation cognitive : elle est ouverte au changement, peu sensibilisée aux biais cognitifs et à la désinformation, et présente une forte activité sur les réseaux sociaux. En influençant subtilement ses perceptions à travers un contenu ciblé, émotionnellement suggestif ou idéologiquement orienté, il est possible d'affecter son évaluation de l'information. Cela pourrait l'amener à favoriser inconsciemment certaines informations, à en reléguer d'autres ou à formuler des recommandations biaisées de manière favorable à l'attaquant, influençant la décision qui en découle.

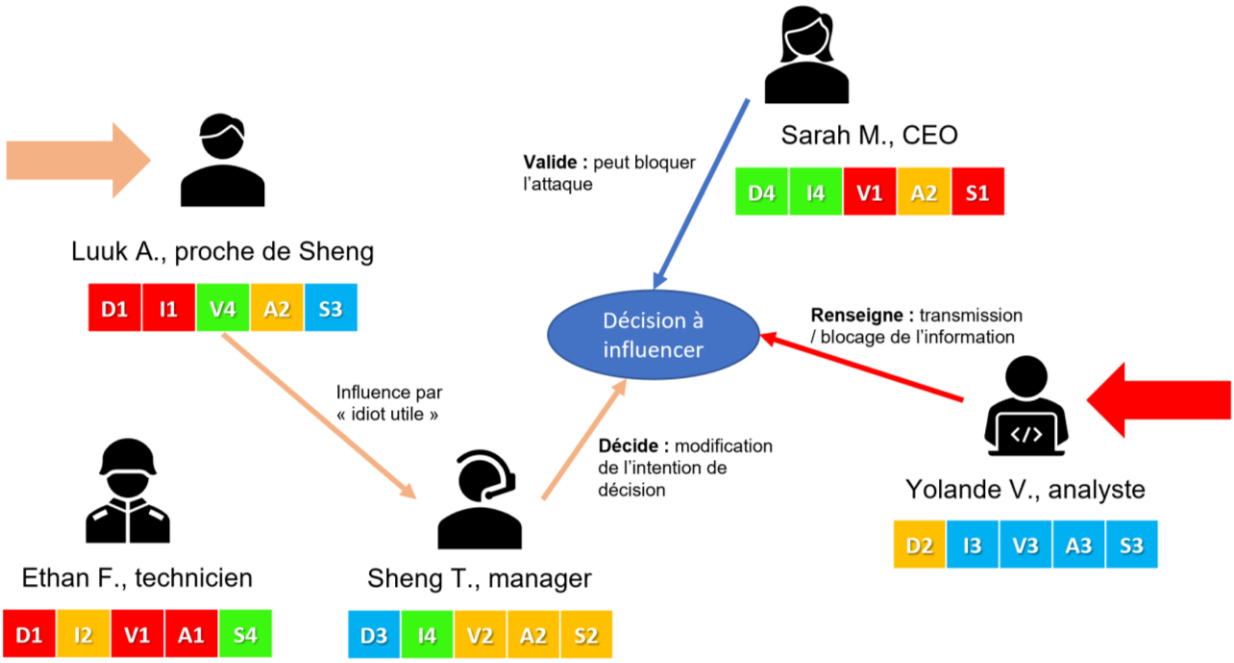

Figure 65 : Exemple d'attaque de la décision possible dans une organisation fictive en s'appuyant sur le modèle DIVAS (en rouge le chemin d'influence idéal, en orange le chemin d'influence alternatif)

Une stratégie alternative ou complémentaire (en orange sur la Figure 65) consiste à instrumentaliser Luuk, proche du manager Sheng. Déjà favorable à la cause de l'attaquant et peu conscient de ses propres vulnérabilités cognitives, il pourrait servir *« d'idiot utile »* pour influencer Sheng sans déclencher ses réflexes de méfiance. En capitalisant sur leur proximité personnelle et sur un discours émotionnel ou narratif inséré dans leurs échanges privés, il est possible d'éroder progressivement les certitudes du

manager, d'introduire des zones de doute, voire de faire évoluer certaines de ses positions sans confrontation directe. Il serait ainsi envisageable d'influencer à plus long terme sa prise de décision.

Cependant, il est à noter que Sarah, en position de validation, peut bloquer ou corriger la décision ainsi influencée. Elle reste donc une cible importante. Si elle ne peut pas être influencée directement dans un délai raisonnable, il reste possible de la décrédibiliser ou de la déstabiliser, par exemple en utilisant les failles cognitives d'Ethan.

Cas général :

Nous pouvons voir ici que les angles d'approche doivent être évalués au cas par cas et nécessitent une analyse qualitative fine.

Le modèle DIVAS doit être utilisé avant tout comme outil d'aide à l'analyse de la situation à un instant donné. La représentation d'une organisation peut évoluer dans le temps : départs ou arrivées de certaines personnes, changements dans les relations hiérarchiques ou informelles, mais aussi changements d'attitude des individus sur certaines caractéristiques évaluées par le modèle, que ce soit dû aux actions de l'agresseur ou à des éléments externes. Il est donc nécessaire de le refaire à intervalles réguliers afin d'observer et prendre en compte les changements qui peuvent survenir.

5.6.2 Quels critères et combinaisons de critères sont les plus importants ?

Nous identifions deux regroupements de critères (Tableau 33) : Axe 1) le rôle dans le processus de prise de décision et la capacité d'influence d'un côté (influence potentielle de l'individu), et Axe 2) la volonté de coopération avec l'initiateur de l'attaque, adaptation au changement et sensibilité à la désinformation et aux failles cognitives de l'autre (vulnérabilités potentielles de l'individu).

Tableau 33 : Regroupement des critères et leur importance

<table>
<tr><td>D : Rôle dans le processus de prise de décision</td><td rowspan="2">Ce groupement de critères détermine l'impact qu'aura la personne si elle est effectivement influencée.

Il est important que l'individu ciblé ait un statut vert ou bleu dans au moins un de ces critères.

Il ne faut cependant pas négliger les personnes ayant une position stratégique dans le réseau ou les cas particuliers (voir les exemples dans la partie 5.5 Modèle obtenu).
Il est intéressant d'évaluer le potentiel d'évolution d'un individu : une personne qui occupe un poste éloigné de la décision qui intéresse l'attaquant, pourrait intégrer plus tard un poste plus stratégique, il pourrait donc être intéressant de cibler cette personne par anticipation.</td></tr>
<tr><td>I : Capacité d'influence</td></tr>
<tr><td>V : Volonté de coopération avec l'attaquant</td><td rowspan="3">Ce groupement de critères détermine la facilité avec laquelle une personne peut être ralliée à la cause de l'attaquant.

Il est donc plutôt important que l'individu ciblé ait un statut vert ou bleu dans au moins un de ces critères : soit il est déjà plutôt favorable à l'attaquant (V), soit il peut être convaincu (A) ou influencé (S) facilement.</td></tr>
<tr><td>A : Disposition à s'adapter au changement</td></tr>
<tr><td>S : Sensibilité à la désinformation et aux failles cognitives</td></tr>
</table>

Notre choix de regroupement repose sur le fait que les individus les plus pertinents et faciles à cibler sont ceux qui présentent au moins un statut bleu ou vert dans un critère dans chacun de ces deux axes.

Le nombre de statuts verts ou bleus minimum pour choisir une personne comme étant une cible privilégiée dépend à la fois de l'analyse qui est faite et du schéma de l'organisation cible, et de l'objectif de l'action. Il est recommandé de viser des personnes ayant un maximum de statuts verts et bleus.

Néanmoins, comme nous l'avons vu plus haut, certains individus peuvent être pris pour cibles même s'ils sont catégorisés comme Saboteurs (rouge) ou Opportunistes (orange) dans le critère des Valeurs, afin de polariser l'opinion encore davantage. De même pour le critère Décision, une personne qui n'intervient pas directement dans la chaîne de décision mais qui Estime (orange) l'information reste très pertinente à cibler, car elle peut perturber l'information de manière directe en influençant sa perception par d'autres membres de l'organisation. Au contraire, un individu qui n'a aucun rôle dans le processus de décision ou aucun contact au sein de l'organisation ciblée peut généralement être directement écarté, quel que soit son statut dans les autres critères.

5.6.3 Stratégies d'attaque en fonction du type de cible

Nous avons déjà mentionné quelques exemples d'attaques potentielles. Par exemple, la cible peut être approchée via les réseaux sociaux ou en direct, dans un cadre privé ou professionnel, avec de la désinformation ou simplement des informations vraies qui ont été bien choisies pour la surprendre ou la déstabiliser, en s'appuyant sur son histoire personnelle ou ses failles cognitives et émotionnelles par exemple.

Il est possible d'utiliser des méthodes d'ingénierie sociale (Mitnick & Simon, 2002 ; David & Bode-Asa, 2023). Elles exploitent la tendance humaine naturelle à faire confiance et d'autres failles psychologiques et sociales (Salahdine & Kaabouch, 2019).

Un autre outil est le kompromat, littéralement *« dossier compromettant »*, une méthode venue de la Russie qui consiste à détruire la réputation d'une personne ou faire pression sur elle en la menaçant de diffuser des informations compromettantes la concernant, que l'on détient ou que l'on fabrique de toutes pièces (Hübert & Little, 2020 ; Claverie, 2023). Utiliser cette méthode sur ses propres alliés induit de nombreux risques : fuite accidentelle des informations compromettantes, risque d'être vu comme une organisation composée essentiellement de membres corrompus, etc. Par ailleurs, un adversaire peut révéler lui-même son kompromat pour qu'il ne puisse plus être utilisé pour l'intimider (Hübert & Little, 2020). De même, cette méthode peut avoir des conséquences importantes sur l'individu ciblé, sa santé et son environnement (insomnie, destruction de l'équilibre familial, financier, etc.) (Claverie, 2023). L'usage du Kompromat apparait peu conforme aux usages des démocraties. Au-delà des considérations éthiques, il faut être conscients de son existence pour s'en protéger. La défense peut passer par exemple par une transparence proactive, qui réduit les leviers de chantage (Nalepa & Sonin, 2020 ; Rid, 2020), ou encore par une cybersécurité avancée, à la fois personnelle et institutionnelle, afin de réduire le risque de vol d'informations sensibles.

5.6.4 Outils pour la qualification des cibles potentielles et l'identification de leurs vulnérabilités

Comme cela a été exposé dans les paragraphes précédents, de nombreuses sources peuvent être exploitées pour obtenir des renseignements sur les cibles potentielles et les qualifier selon les différents

critères que nous avons identifiés. Par exemple, via un renseignement humain et technique, en exploitant les informations disponibles sur les réseaux sociaux (messages postés, personnes suivies, forme du réseau de contacts, ce que d'autres disent de la cible etc.), ou sur certaines plateformes offrant un tracé GPS (comme les outils de suivi de santé ou de sport), en récoltant les informations disponibles en sources ouvertes sur l'organisation, etc. (Hung & Hung, 2022 ; Hristova et al., 2014 ; Voelker et al., 2011 ; Williams et al., 2015).

Nous avons également mentionné l'existence de questionnaires qui permettent d'évaluer certains critères (Oreg, 2003 ; Maertens et al., 2023 ; Piksa et al., 2022). Bien sûr, il est difficile de faire passer des questionnaires aux membres du C2 ou organisation adverse visés. L'analyse nécessaire pour construire un schéma du C2 suivant le modèle DIVAS repose essentiellement sur le renseignement et l'interprétation individuelle de chaque personne impliquée dans l'organisation ou le processus de décision cible. Cependant, ces questionnaires peuvent déjà nous aider à évaluer nos propres failles et éventuellement à déduire, à l'issue d'un travail de recherche, d'autres caractéristiques plus facilement observables des cibles potentielles. Les questionnaires présentent d'autres défauts pour évaluer correctement le profil psychologique d'une personne, tels que le biais de désirabilité sociale qui peut fausser les résultats, ou encore le fait qu'ils mesurent souvent des opinions conscientes et explicites (Smith, 2023 ; Choi & Pak, 2004).

En revanche, les données massives en sources ouvertes (Big Data) sont plus difficiles à recueillir et analyser, mais peuvent permettre de prédire la personnalité, l'humeur, les émotions ou encore l'orientation sexuelle d'une personne à partir de son activité sur les réseaux sociaux, la navigation web, les musiques écoutées, les capteurs des smartphones etc. (Kosinski, 2019). De nombreux renseignements peuvent également être obtenus en sources ouvertes, notamment grâce aux nombreuses données exposées sur les réseaux sociaux, qui permettent d'identifier les biais exploitables d'une personne et prendre contact avec elle tout en lui donnant l'impression de très bien la connaître (Jolicard & Gardin, 2024). L'utilisation des big data et du machine learning pour l'analyse des préférences et du comportement à des fins de microciblage est aussi une piste prometteuse (Ehrhard, 2019), par exemple pour reconnaître les personnes les plus vulnérables émotionnellement. Il est possible de s'inspirer pour cela des méthodes de l'analyse du public cible (voir paragraphe *5.3 Modèles de ciblage documentés dans la littérature*) utilisées par de nombreuses plateformes pour la diffusion de publicités ciblées ou la suggestion de contenus (Debidour & Pelletier, 2024 ; Bogacki et al., 2024). L'exemple du scandale Cambridge Analytica (Isaak & Hanna, 2018) montre qu'il est possible de déduire de nombreuses informations sur les cibles potentielles à partir de leur utilisation des réseaux sociaux notamment.

Nous soulignons donc l'importance d'utiliser des outils d'analyse dédiés, capables de traiter l'importante masse de données disponibles et d'en tirer des informations utiles pour identifier les points de vulnérabilité d'une organisation, et de déterminer les moyens les plus susceptibles d'être utilisés par un attaquant en fonction de sa cible et de ses intentions. Il s'agit justement de l'objectif que nous souhaitons atteindre par le biais de cette thèse, en étudiant l'adaptation du logiciel ANTICIPE[11] dans un cadre de guerre cognitive.

[11] https://www.youtube.com/watch?v=A2ZAHrT3UwM

5.7 Conclusion de l'étude prospective

Cette étude prospective avait pour objectif de créer une grille d'analyse pour catégoriser les critères des cibles potentielles d'attaques de guerre cognitive. Dans notre démarche de thèse, cette étape constitue un des outils nécessaires que nous rassemblons pour construire ou adapter un système d'aide à la détection d'attaques de guerre cognitive ou à l'établissement de contremesures.

Ainsi, nous avons pu identifier les éléments qui permettent de mieux distinguer les cibles potentielles, en s'intéressant à la fois à l'impact (influence sur d'autres, rôle dans un processus décisionnel) et aux vulnérabilités (valeurs, disposition à s'adapter au changement, failles cognitives) de la personne ou la structure visée.

Nous avons identifié différents moyens pour qualifier un individu suivant les critères proposés : l'investigation via les données disponibles sur les réseaux sociaux, l'étude des réseaux professionnels, extra-professionnels et privés, des questionnaires… Ces éléments doivent aussi nous apporter un éclairage pour mieux se protéger des attaques de guerre cognitive, en identifiant nos propres points de vulnérabilité et comment les protéger. Il faudra également prendre garde aux informations qui peuvent être disponibles sur nos agents afin qu'un tel modèle ne puisse pas être utilisé contre nous. Cela passe notamment par la protection de leurs données via la cybersécurité et une conscience des risques cyber, à la fois institutionnelle et individuelle. Pour éviter la création de kompromat sur des agents, ils peuvent être sensibilisés aux comportements utilisables contre eux et contre leur organisation. Nous soulignons également l'importance d'être attentifs à la personnalité et à l'intégrité des individus ayant un rôle important dans la prise de décision ou une capacité d'influence importante sur d'autres agents. Par ailleurs, il est primordial d'organiser un entraînement à la résilience cognitive de toute personne intervenant dans la chaîne de prise de décision. Cet entraînement doit combiner la connaissance des biais cognitifs, la formation à l'esprit critique, la gestion du stress décisionnel et la coopération organisationnelle ; il doit ainsi viser à renforcer la capacité à détecter, comprendre et contrer les tentatives d'influence (Ketelaars et al., 2024 ; Weick & Sutcliffe, 2015).

L'identification des cibles potentielles pourrait se faire via un outil dédié qui soit capable de traiter la masse de données disponibles et d'en tirer des informations pertinentes sur l'importance de l'influence d'une personne au sein d'une organisation ainsi que ses vulnérabilités potentielles. Nous proposons ainsi d'adapter le système ANTICIPE à la guerre cognitive, tel qu'illustré avec Postdare dans le *Chapitre 4* -.

Cependant, il est important de conserver en mémoire les règles d'éthique et de respect de la vie privée des personnes dont les caractéristiques seraient scrutées par un tel outil.

Cette grille d'analyse peut être appliquée à un C2 de forces armées mais peut également concerner une entreprise ou toute autre organisation critique susceptible d'être ciblée par une attaque de guerre cognitive.

Conclusion générale

Dans un monde où l'information et les connaissances sont devenues à la fois une arme et une cible, la guerre cognitive s'impose comme un nouveau champ de confrontation stratégique. Invisible, polymorphe, transnationale, elle infiltre les esprits, manipule les perceptions et altère les décisions sans que ses victimes n'en aient nécessairement conscience. Dans ce contexte, les décideurs, qu'ils soient militaires, politiques, économiques ou institutionnels, représentent des cibles privilégiées en raison de leur pouvoir d'influence et de la portée de leurs décisions.

Cette thèse avait pour objectif de poser les bases d'un système d'aide à la décision dans un contexte de guerre cognitive. À cette fin, nous avons développé une démarche pluridisciplinaire combinant une revue de littérature, deux expérimentations et une réflexion prospective sur la qualification des cibles potentielles. Nous soulignons par là l'intérêt de croiser l'ingénierie cognitique, les sciences de l'information, la psychologie sociale et les études stratégiques pour mieux comprendre et prévenir les phénomènes d'influence.

La première partie a permis d'étudier les fondements conceptuels de la guerre cognitive, d'en analyser les outils, les acteurs, les modes opératoires et les enjeux. Ce mode de conflit discret, capable de prendre pour cible l'esprit de tout humain connecté au monde numérique, doit être adressé pour assurer la sécurité cognitive et la stabilité des démocraties et de leurs citoyens (voir les bilans pages 26, 51 et 57 – *Chapitre 1*). Nous avons également mis en évidence les vulnérabilités des processus cognitifs humains, notamment celles des décideurs, face à des stratégies d'influence ciblées. Nous avons montré la nécessité de construire des systèmes d'aide à la prise de conscience de situation et de support à la décision pour protéger les décideurs des agressions relevant de la guerre cognitive (bilans pages 62, 68 et 73 – *Chapitre 2*).

La deuxième partie a présenté une démarche empirique permettant de poser les fondements de différentes dimensions du système visé : des outils permettant de mettre en place des contremesures d'influence (première expérimentation), et le processus permettant de construire, améliorer et utiliser l'arbre de détection d'informations critiques nécessaire pour détecter les signaux faibles et qualifier les situations (deuxième expérimentation).

Ainsi, la première étude (*Chapitre 3*) a exploré les leviers cognitifs de l'influence à travers une expérimentation portant sur la crédibilité perçue des messages textuels. Elle a permis d'identifier douze facteurs susceptibles d'augmenter l'impact persuasif d'un message, parmi lesquels les plus influents se sont avérés être : la présence de détails, d'exemples, de sources et de répétitions dans le message. Nous avons également trouvé des différences entre certains profils, les participants les plus diplômés tendant à accorder plus de confiance que les autres à la présence de sources ; les participants les plus sensibles à la désinformation accordaient plus de confiance à la présence de chiffres que les autres (voir *3.8 Conclusion de l'expérimentation 1*). Ces résultats fournissent des pistes précieuses pour la conception d'outils de détection et de prévention des tentatives de manipulation, mais aussi pour l'amélioration de messages destinés à mener une contre-influence.

La deuxième expérimentation (*Chapitre 4*) a pris l'exemple d'un jeu de société semi-collaboratif offrant des possibilités de coopération comme de compétition. Elle a permis d'établir le potentiel de l'adaptation du système ANTICIPE pour détecter les stratégies dominantes en temps réel à partir de données récoltées automatiquement et pour proposer des réponses adaptées, sous forme d'actions d'influence. Cette étude a permis d'expérimenter les différentes phases de conception et d'ajustement du système pour l'utiliser

dans de nouvelles situations, en vue de son application à la guerre cognitive (voir *4.4 Conclusion de l'expérimentation 2*).

Dans la troisième partie, une réflexion prospective (*Chapitre 5*) a conduit à la proposition d'un modèle de qualification des cibles de guerre cognitive, articulé autour de cinq dimensions clés : Décision, Influence, Volonté de coopération, Adaptation et Sensibilité à la désinformation et aux failles cognitives. Ce modèle propose une base méthodologique pour évaluer la pertinence d'exploiter une cible en vue d'actions cognitives dirigées, en fonction de son potentiel d'influence et de sa vulnérabilité (voir *5.7 Conclusion de l'étude prospective*).

L'ensemble de ces travaux contribue à la construction d'un socle méthodologique pour le développement d'un système technologique d'aide à la conscience de situation et à la prise de décision dans un contexte de guerre cognitive. Un tel système aurait pour ambition de mieux comprendre les mécanismes d'influence, détecter les attaques cognitives, affiner le ciblage, choisir des interventions offensives ou défensives adaptées, mais aussi renforcer la résilience décisionnelle des acteurs clés (Figure 66).

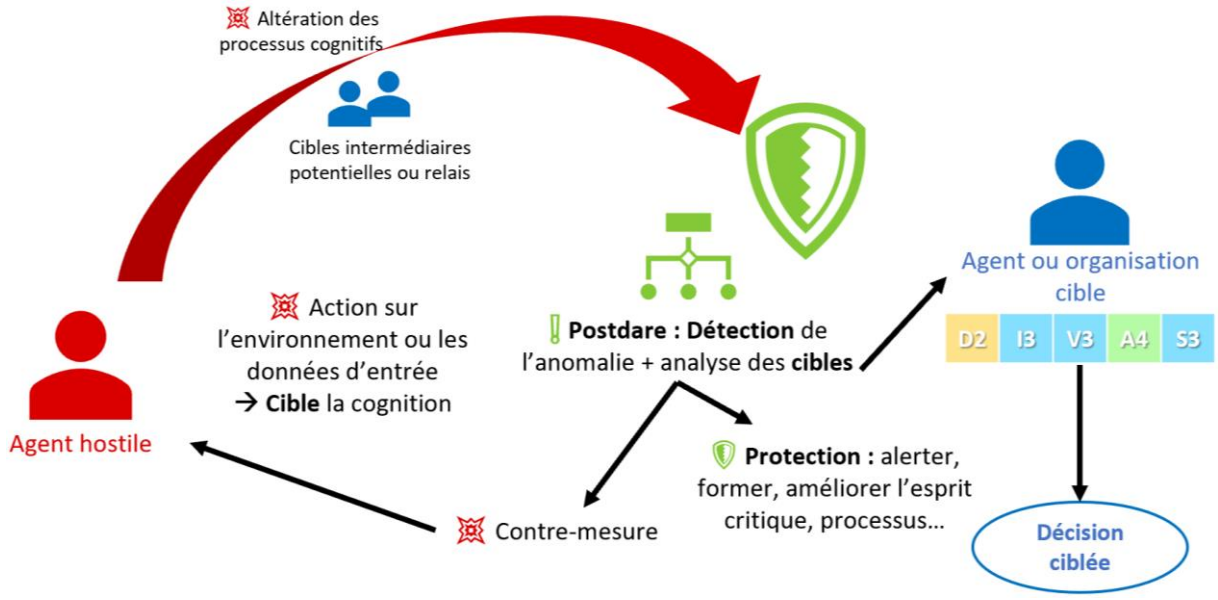

Figure 66 : Intervention défensive de l'outil ANTICIPE / Postdare afin de protéger l'agent ou organisation cible d'une action de guerre cognitive entreprise par un agent hostile

Ce travail ouvre également des perspectives opérationnelles : l'adaptation d'ANTICIPE au contexte de la guerre cognitive (sous forme du logiciel Postdare) constitue une première étape vers la mise en œuvre d'outils de veille, de détection et de protection face aux menaces cognitives. À terme, ce type d'architecture pourrait être utilisé dans des cellules de cybersécurité, de renseignement ou de communication stratégique, afin de soutenir la prise de décision et d'améliorer la vigilance cognitive collective.

Comme tout travail exploratoire, cette recherche comporte certaines limites. Les facteurs d'influence étudiés dans la première expérimentation ont été restreints aux messages textuels, alors qu'une analyse exhaustive devrait inclure d'autres formes médiatiques telles que l'image, la vidéo ou les environnements interactifs, dont les effets cognitifs peuvent différer sensiblement. Dans la deuxième expérimentation, le temps disponible a conduit à tester principalement des situations d'influence directe et à court terme, en laboratoire. Des formes plus fines d'influence (indirectes, différées ou cumulatives dans le temps) mériteraient d'être étudiées. Le modèle DIVAS, bien que prometteur, demeure théorique et nécessitera des validations empiriques pour confirmer sa robustesse et sa transférabilité à d'autres contextes. Enfin,

ce travail s'inscrit dans une démarche prospective et de long terme ; notamment, nous n'avons pas exploré la détection de formes nouvelles de guerre cognitive non connues par le système. Or, les moyens de la guerre cognitive évolueront sans doute avec les technologies et les contextes sociopolitiques. Nous considérons qu'il existera probablement une nécessité de supervision humaine du système ANTICIPE / Postdare imaginé, notamment pour assurer une veille, une interprétation contextuelle et une mise à jour continue des indicateurs de détection. À terme, chaque schéma d'attaque identifié et validé manuellement pourrait être intégré dans le système, permettant ainsi d'automatiser la reconnaissance de situations similaires lors d'occurrences futures.

Au-delà des résultats obtenus, cette thèse invite donc à poursuivre les recherches dans plusieurs directions :

- croiser des regards d'experts pour déterminer les indices qui permettent de la détecter, afin de concevoir un arbre de détection d'informations critiques adapté à la guerre cognitive ;
- intégrer des données multimodales ;
- proposer des actions d'influence adaptées à la situation ;
- adapter et évaluer le système dans des conditions plus proches de la réalité (simulation, wargame, exercice militaire) voire en conditions réelles ;
- consolider et évaluer le modèle théorique DIVAS ;
- explorer les questions éthiques posées par l'usage de telles technologies.

Ces travaux devront s'accompagner d'une réflexion éthique, afin de garantir que ces outils restent au service de la protection et de la résilience, et non de la manipulation. Il apparaît nécessaire de définir des cadres de gouvernance scientifique et déontologique fondés sur la transparence, la traçabilité et la supervision humaine. La lutte contre la guerre cognitive doit reposer sur une coopération entre chercheurs, institutions et acteurs publics, afin de préserver l'intégrité cognitive des sociétés démocratiques.

Face à une guerre qui ne dit pas son nom mais agit sur les fondements mêmes de notre autonomie mentale et décisionnelle, nous soulignons l'importance de renforcer nos défenses cognitives, de concevoir des outils intelligents d'aide à la décision, et de promouvoir une culture de la vigilance et de l'esprit critique. Ce travail constitue une contribution à cet effort collectif.

Bibliographie

Agocs, C. (1997). Institutionalized Resistance to Organizational Change: Denial, Inaction and Repression. *Journal of Business Ethics, 16*, 917-931. https://doi.org/10.1023/A:1017939404578

Akinwumi, R. P., & Olakunle, A. O. (2025). The impact of fake news on public perception of science: An empirical review. *Advance Journal of Linguistics and Mass Communication, 9*(1), 1.

Alaphilippe, A., Gizikis, A., Hanot, C., & Bontcheva, K. (2019). *Automated tackling of disinformation*. Panel for the Future of Science and Technology, European Science-Media Hub. www.sheffield.ac.uk/sites/default/files/2022-04/EPRS_STU2019624278_EN.pdf

Albertini, A., Leloup, D., & Reynaud, F. (2023, novembre 7). Etoiles de David taguées à Paris : La piste d'une opération d'ingérence russe privilégiée. *Le Monde.* www.lemonde.fr/societe/article/2023/11/07/pochoirs-d-etoiles-de-david-a-paris-la-piste-d-une-operation-d-ingerence-russe-privilegiee_6198775_3224.html

Albou, P., Crocq, L., Flusin, J.-P., Fourcade, J., & Rivolier, J. (1990). *Stress et prise de décision*. Dossier N°31, Fondation pour les études de défense nationale.

Allain, P. (2013). La prise de décision : Aspects théoriques, neuro-anatomie et évaluation. *Revue de neuropsychologie, 5*(2), 69-81. https://doi.org/10.1684/nrp.2013.0257

Amalberti, R. (1996). *La conduite de systèmes à risques* (Presses Universitaires de France).

Amblard, H., Bernoux, P., Herreros, G., & Livian, Y.-F. (2005). *Les Nouvelles approches sociologiques des organisations* (3e éd.). Éditions du Seuil.

American Psychological Association. (2023, novembre 29). *What psychological factors make people susceptible to believe and act on misinformation?* www.apa.org/topics/journalism-facts/misinformation-belief-action

Ananthaswamy, A. (2016). *Virtual reality could be an ethical minefield – are we ready?* New Scientist. Consulté le 3 janvier 2023, à l'adresse www.newscientist.com/article/2079601-virtual-reality-could-be-an-ethical-minefield-are-we-ready/

Anaut, M. (2015). *Psychologie de la résilience*. Armand Colin. https://shs.cairn.info/psychologie-de-la-resilience--9782200611798

Anderson, B. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism.*

Anonyme. (XIVe à XVIIe siècle). *Les 36 stratagèmes : Manuel secret de l'art de la guerre*.

Antoniuk, D. (2024). Russian influence operations against Baltic states and Poland having 'significant impact' on society. *The Record Media*. https://therecord.media/russian-influence-operations-baltic-poland-impact

Ariely, D. (2008). *Predictably Irrational: The hidden forces that shape our decisions* (HarperCollins Canada).

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110*(3), 486–498. https://doi.org/10.1037/0033-2909.110.3.486

Arnott, D. (1998). A Taxonomy of Decision Biases. *Monash University - School of Information Management & Systems.*

Aro, J. (2016). The Cyberspace War: Propaganda and Trolling as Warfare Tools. *European View, 15*(1), 121-132. https://doi.org/10.1007/s12290-016-0395-5

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *The Journal of Abnormal and Social Psychology, 59*(2), 177–181. https://doi.org/10.1037/h0047195

Arreguín-Toft, I. (2001). How the Weak Win Wars: A Theory of Asymmetric Conflict. *International Security, 26*(1), 93-128.

Ask, T. F., Lugo, R. G., Sütterlin, S., Canham, M., Hermansen, D., & Knox, B. J. (2023). *The UnCODE System : A neurocentric systems approach for classifying the goals and methods of Cognitive Warfare.* Preprint.

Autellet, E. (2021). Chapitre 3 – Cognitive Warfare - MGA - Contribution du major général des armées (République Française). Dans B. Claverie, B. Prébot et F. Du Cluzel (dir.), *La Guerre Cognitique* (p. 3.1-3.2). CSO

Backes, O., Swab, A. (2019). *Cognitive Warfare—The Russian Threat to Election Integrity in the Baltic States*. Doctoral dissertation, Harvard University.

Badawy, A., Addawood, A., Lerman, K., & Ferrara, E. (2019). Characterizing the 2016 Russian IRA influence campaign. *Social Network Analysis and Mining, 9*(1), 31. https://doi.org/10.1007/s13278-019-0578-6

Badawy, A., & Ferrara, E. (2018). The Rise of Jihadist Propaganda on Social Networks. *Journal of Computational Social Science, 1*(2), 453-470. https://doi.org/10.1007/s42001-018-0015-z

Baezner, M., & Robin, P. (2017). Cyber-conflict between the United States of America and Russia. Dans CSS *Cyberdefense Hotspot Analyses* (Numéro 2) [Report]. ETH Zurich. https://doi.org/10.3929/ethz-b-000169642

Baker, D. F. (2010). Enhancing Group Decision Making: An Exercise to Reduce Shared Information Bias. *Journal of Management Education, 34*(2), 249-279. https://doi.org/10.1177/1052562909343553

Baptista, J. P., & Gradim, A. (2020). Understanding Fake News Consumption: A Review. *Social Sciences, 9*(10), 10. https://doi.org/10.3390/socsci9100185

Baptista, A., Soumet-Leman, C., & Jouvent, R. (2014). Métacognition et dépression : Validation d'une version française du MCQ-30 en population clinique. *European Psychiatry, 29*(S3), 569-569. https://doi.org/10.1016/j.eurpsy.2014.09.251

Baron, J. (2007). *Thinking and Deciding* (4th Edition).

Bartheye, O., & Chaudron, L. (2018). Algebraic Modeling of the Causal Break and Representation of the Decision Process in Contextual Structures. Dans *Computational Context*. CRC Press.

Bateman, J., & Jackson, D. (2024). *Countering Disinformation Effectively: An Evidence-Based Policy Guide.* Carnegie Endowment for International Peace. carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide

Battrawi, B., & Muhtaseb, R. (2013). The Use of Social Networks as a Tool to Increase Interest in Science and Science Literacy: A Case Study of « Creative Minds » Facebook Page. *New Perspectives in Science Education 2.*

Baudel, T., Verbockhaven, M., Cousergue, V., Roy, G., & Laarach, R. (2021). ObjectivAIze: Measuring Performance and Biases in Augmented Business Decision Systems. *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference*, 300-320.

Baughman, J., & Singer, P. W. (2023, octobre 17). *China's social-media attacks are part of a larger 'cognitive warfare' campaign*. Defense One. www.defenseone.com/ideas/2023/10/chinas-social-media-attacks-are-part-larger-cognitive-warfare-campaign/391255/

Baumard, P. (2002). Les limites d'une économie de la guerre cognitive. Dans C. Harbulot & D. Lucas (Eds.), *La guerre cognitive* (p. 35-55). hal-03230319

Beauchamp-Mustafaga, N. (2023). *Chinese Next-Generation Psychological Warfare: The Military Applications of Emerging Technologies and Implications for the United States*. RAND. www.rand.org/pubs/research_reports/RRA853-1.html

Beauvois, J.-L., & Joule, R.-V. (2020). *La dissonance cognitive : Approche socio-cognitive*. Armand Colin

Bebber, R. J. (2024). *Cognitive Competition, Conflict, and War: An Ontological Approach*. Hudson Institute. www.hudson.org/defense-strategy/cognitive-competition-conflict-war-ontological-approach-robert-jake-bebber

Becker, J.-J. (1978). L'opinion publique française et les débuts de la guerre de 1914 (printemps-Automne 1914). *Le Mouvement social, 104*, 63-73. https://doi.org/10.2307/3778104

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin, 144*(1), 77-110. https://doi.org/10.1037/bul0000130

Begou, N., Vinoy, J., Duda, A., & Korczyński, M. (2023). Exploring the Dark Side of AI: Advanced Phishing Attack Design and Deployment Using ChatGPT. *2023 IEEE Conference on Communications and Network Security*, 1-6. https://doi.org/10.1109/CNS59707.2023.10288940

Bell, M. A., Facci, E. L., Nayeem, R. V. (2005). Cognitive Tunneling, Aircraft-Pilot Coupling Design Issues and Scenario Interpretation Under Stress in Recent Airline Accidents. *2005 International Symposium on Aviation Psychology*, 45-49.

Béna, J., Carreras, O., & Terrier, P. (2019). L'effet de vérité induit par la répétition : revue critique de l'hypothèse de familiarité. *L'Année psychologique, 119*(3), 397-425.

Bennabi, D., & Haffen, E. (2018). Transcranial Direct Current Stimulation (tDCS): A Promising Treatment for Major Depressive Disorder? *Brain Sciences, 8*(5), 5. https://doi.org/10.3390/brainsci8050081

Berghel, H. (2018). Malice Domestic: The Cambridge Analytica Dystopia. *Computer, 51*(5), 84-89. https://doi.org/10.1109/MC.2018.2381135

Bernal, A., Carter, C., Singh, I., Cao, K., Madreperla, O. (2020). *Cognitive Warfare: An Attack on Truth and Thought.* NATO and Johns Hopkins University: Baltimore MD, USA.

Bernays, E. (1928). *Propaganda*. Horace Liveright.

Bertrand, B. (2023a). Désinformation : Quels enjeux ? Quels effets systémiques ? In *Lutte contre les manipulations de l'information, regards croisés des experts du domaine* (Pôle d'excellence cyber, Vol. 1).

Bertrand, B. (2023b). Lutte contre la désinformation. Quelle régulation juridique ? In *Lutte contre les manipulations de l'information, regards croisés des experts du domaine* (Pôle d'excellence cyber, Vol. 1).

Bertrand, B., Nocetti, J., Moreau, J., Riant, J.-P., Puren, M., Laizé, A., Paquelet, S., Labouré, E., Bischetti, P., Mabille, C., Courtois, B., Delavallade, T., Pahl, M.-O., Faviere, F., Turgis, S., & Bresson, E. (2023). *Lutte contre les manipulations de l'information, regards croisés des experts du domaine* (Pôle d'excellence cyber).

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences, 41*, 15-21. https://doi.org/10.1016/j.cobeha.2021.02.015

Black, I., & Morris, B. (1991). *Israel's Secret Wars: A History of Israel's Intelligence Services*. Grove Press.

Blanco, F. (2017). Cognitive Bias. Dans J. Vonk & T. Shackelford (Éds.), *Encyclopedia of Animal Cognition and Behavior* (p. 1-7). Springer International Publishing. https://doi.org/10.1007/978-3-319-47829-6_1244-1

Bobric, G.-D. (2021). The Overton Window: A Tool for Information Warfare. Dans J. Lopez, K. Perumalla et A. Siraj, I*CCWS 2021 16th International Conference on Cyber Warfare and Security* (p. 20-27). Academic Conferences Limited.

Bogacki, S., Smutek, T., Chmielowska-Marmucka, A., Rymarczyk, P., & Rutkowski, M. (2024). Advanced methods for target audience identification: Enhancing marketing strategies through machine learning and data analytics. *Journal of Modern Science, 57*, 417. https://doi.org/10.13166/jms/191180

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes, 101*(2), 127-151.

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy, 3*, e32. https://doi.org/10.1017/dap.2021.20

Borgonovo, F. (2022). Strategies, disinformation techniques and cognitive warfare of jihadist organisations. *The CoESPU MAGAZINE - the Online Journal of Stability Policing - Advanced Studies, 1*(1), 41. https://doi.org/10.32048/Coespumagazine4.22.10

Botelho, L. M., & Coelho, H. (1996). Information Processing, Motivation and Decision Making. Dans P. Ein-Dor (Éd.), *Artificial Intelligence in Economics and Management: Edited Proceedings on the Fourth International Workshop: AIEM4* (p. 233-250). Springer US. https://doi.org/10.1007/978-1-4613-1427-1_16

Botti, S., & lyengar, S. S. (2004). The Psychological Pleasure and Pain of Choosing: When People Prefer Choosing at the Cost of Subsequent Outcome Satisfaction. *Journal of Personality and Social Psychology, 87*(3), 312-326. https://doi.org/10.1037/0022-3514.87.3.312

Boyer, B. (2024, avril 28). *S'armer pour la guerre cognitive : Le modèle DIMA*. M82 Project. https://m82-project.org:443/articles/dima/dima/

Box, G. E. (1953). Non-normality and tests on variances. *Biometrika, 40*(3/4), 318-335.

Brady, H. E., & Kent, T. B. (2022). Fifty Years of Declining Confidence & Increasing Polarization in Trust in American Institutions. *Daedalus, 151*(4), 43-66. https://doi.org/10.1162/daed_a_01943

Brehm J. W. (1956). Postdecision changes in the desirability of alternatives. *Journal of abnormal psychology, 52*(3), 384–389. https://doi.org/10.1037/h0041006

Brenner, P. (1990). Cuba and the missile crisis. *Journal of Latin American Studies, 22*(1-2), 115-142. https://doi.org/10.1017/S0022216X00015133

Bridges, W. (1991). *Managing Transitions: Making The Most Of Change*. Da Capo Press.

Brittain-Hale, A. (2023). Clausewitzian Theory Of War in The Age of Cognitive Warfare. *The Defence Horizon Journal*, 12.

Brunyé, T. T., Brou, R., Doty, T. J., Gregory, F. D., Hussey, E. K., Lieberman, H. R., Loverro, K. L., Mezzacappa, E. S., Neumeier, W. H., Patton, D. J., Soares, J. W., Thomas, T. P., Yu, A. B. (2020). A Review of US Army Research Contributing to Cognitive Enhancement in Military Contexts. *Journal of Cognitive Enhancement, 4*(4), 453-468. https://doi.org/10.1007/s41465-020-00167-3

Buch, B., & Mitchell, K. (2013). The Active Denial System: Obstacles and promise. *Convention on Certain Conventional Weapons.*

Buchler, N. (2021). Chapitre 6 – Maturité technique des systèmes cognitifs des réseaux humains. Dans B. Claverie, B. Prébot et F. Du Cluzel (dir.), *La Guerre Cognitique* (p. 6.1-6.13). CSO

Bühlmann, C. (2009). Asymmetric strategies : A concept to better understand modern conflicts? *Military Power Revue Der Schweizer Armee, 2*, 8-21.

Bulinge, F. (2010). Renseignement militaire : une approche épistémologique. *Revue internationale d'intelligence économique,* 2, 209-232. www.cairn.info/revue--2010-2-page-209.htm

Burt, R. S. (2014). Structural Holes. Dans *Social Stratification* (4e éd.). Routledge.

Cacioppo, J. T., & Hawkley, L. C. (2009). Perceived social isolation and cognition. *Trends in Cognitive Sciences, 13*(10), 447-454. https://doi.org/10.1016/j.tics.2009.06.005

Caillé, A. (2016). Pouvoir, domination, charisme et leadership. *Revue du MAUSS, 47*(1), 305-319. https://doi.org/10.3917/rdm.047.0305

Caire, J., & Conchon, S. (2018). Influence 2.0 : Comprendre les opérations d'influence dans un monde hyperconnecté. *Congrès Lambda Mu 21, « Maîtrise des risques et transformation numérique : opportunités et menaces »*, Reims, France. https://hal.science/hal-02071177/

Calo, R. (2013). Digital Market Manipulation. *George Washington Law Review, 82*, 995.

Camilleri, M. A. (2018). Market Segmentation, Targeting and Positioning. Dans *Travel Marketing, Tourism Economics and the Airline Product* (Chapter 4, p. 69-83).

Campos, N. F., & Giovannoni, F. (2006). Lobbying, corruption and political influence. *Public Choice, 131*(1), 1-21. https://doi.org/10.1007/s11127-006-9102-4

Cao, K., Glaister, S., Pena, A., Rhee, D., Rong, W., Rovalino, A., Bishop, S., Khanna, R., & Singh Saini, J. (2021). Countering cognitive warfare: Awareness and resilience. *NATO Innovation Hub*. www.nato.int/docu/review/articles/2021/05/20/countering-cognitive-warfare-awareness-and-resilience

Castel, P., & Chessel, M.-E. (2024). *À la recherche de la décision* (Presses Universitaires du Septentrion). https://www.septentrion.com/fr/book/?GCOI=27574100119740

Charon, P., Jeangène Vilmer, J.-B. (2021). *Les Opérations d'influence chinoises : Un moment machiavélien* (2e édition). Rapport de l'Institut de recherche stratégique de l'École militaire (IRSEM), Paris, ministère des Armées.

Charzat, X. (2024). Décider malgré la perplexité : Le cerveau du chef comme champ de bataille. *Cahiers de Conflits, 3,* 11-24. https://doi.org/10.3917/cdc.243.0011

Chaudron, L., Erceau, J., & Trousse, B. (1993). Cooperative decisions and actions in multi-agent worlds. *Proceedings of IEEE Systems Man and Cybernetics Conference - SMC, 3*, 626-631 https://doi.org/10.1109/ICSMC.1993.385085

Chauvancy, R. (2021). L'ingénierie cognitive, arme de guerre. *Revue internationale d'intelligence économique, 13*(2), 11-22.

Chavalarias, D. (2024). Minuit moins dix à l'horloge de Poutine. *Pre-print de l'Institut des Systèmes Complexes de Paris Île-de-France.*

Chemero, A. (2023). Abduction and Deduction in Dynamical Cognitive Science. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12692

Cherré, B., Laarraf, Z., & Yanat, Z. (2014). Dissonance éthique : Forme de souffrance par la perte de sens au travail. *Recherches en Sciences de Gestion, 100*(1), 143-172. https://doi.org/10.3917/resg.100.0143

Cheskin, A. (2024). Identity and integration of Russian speakers in the Baltic States: A framework for analysis. Dans *Diasporas and Transportation of Homeland Conflicts* (pp. 156-177). Routledge.

Choi, B. C. K., & Pak, A. W. P. (2004). A Catalog of Biases in Questionnaires. *Preventing Chronic Disease, 2*(1), A13.

Clausewitz, C. von. (1832). *Vom Kriege (De la guerre).*

Claverie, B. (2021). Chapitre 4 - Qu'est-ce que la cognition et comment en faire l'un des moyens de la guerre ? Dans B. Claverie, B. Prébot et F. du Cluzel (dir.), *La Guerre Cognitique* (p. 4.1-4.20). CSO

Claverie, B. (2023). Les opérations d'influence psychologiques russes et la Maskirovka comme état d'esprit. *ISTE Open Science*. www.openscience.fr/Les-operations-d-influence-psychologiques-russes-et-la-Maskirovka-comme-etat-d

Claverie, B., & Desclaux, G. (2022). C2 - Command and Control : A System of Systems to Control Complexity. *American Journal of Management, 22*(2). https://doi.org/10.33423/ajm.v22i2.5381

Claverie, B., du Cluzel, F. (2021). Chapitre 1 - Le Cognitive Warfare et l'avènement du concept de « Guerre Cognitique ». Dans B. Claverie, B. Prébot et F. du Cluzel (dir.), *La Guerre Cognitique* (p. 1.1-1.8). CSO

Claverie, B., Prébot, B., & Du Cluzel, F. (2021). Cognitive Warfare : La guerre cognitique (CSO). https://aorcompiegne.fr/cognitive-warfare-la-guerre-cognitique

Coccia, M. (2020). Critical decisions in crisis management: Rational strategies of decision making. *Journal of Economics Library, 7*(2), 81-96.

Cocron, A., & Aronhime, L. (2022). *Cognitive Warfare: What's Next?* NATO Cognitive Warfare Workshop - USMA West Point.

Colon, D. (2021). *Les maîtres de la manipulation : Un siècle de persuasion de masse.* Tallandier.

Connolly, M., Hawkshaw, M. J., & Sataloff, R. T. (2024). Havana syndrome: Overview for otolaryngologists. *American Journal of Otolaryngology, 45*(4), 104332. https://doi.org/10.1016/j.amjoto.2024.104332

Cornish, P., Livingstone, D., Clemente, D., & Yorke, C. (2010). *On Cyber Warfare*. The Royal Institute of International Affairs - Chatham House. www.chathamhouse.org/sites/default/files/public/Research/International%20Security/r1110_cyberwarfare.pdf

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science, 385*(6714). https://doi.org/10.1126/science.adq1814

Courbet, D., & Benoit, D. (2013). Neurosciences au service de la communication commerciale : Manipulation et éthique. Une critique du neuromarketing. *Études de communication, 40*. https://doi.org/10.4000/edc.5091

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM, 63*(10), 72-83. https://doi.org/10.1145/3409116

Cristofaro, M. (2017). Reducing biases of decision-making processes in complex organizations. *Management Research Review, 40*(3), 270-291. https://doi.org/10.1108/MRR-03-2016-0054

Crozier, M., & Friedberg, E. (1977). *L'Acteur et le Système : Les contraintes de l'action collective*. Éditions du Seuil.

Cvitkovic, P. (2024). Désinformation : Comprendre, détecter et agir. Dans *Lutte contre les manipulations de l'information—Regards croisés de spécialistes et d'acteurs du domaine* (Pôle d'Excellence Cyber, Vol. 2). www.pole-excellence-cyber.org/evenements/lutte-contre-les-manipulations-de-linformation-decouvrez-le-tome-2/

Danciu, V. (2014). Manipulative marketing: Persuasion and manipulation of the consumer through advertising. *Theoretical and Applied Economics, 21*(2), 19-34.

Danylov, O. (2022). *The unique Ukrainian situational awareness system Delta was presented at the annual NATO event*. Mezha. Consulté le 5 décembre 2022, à l'adresse https://mezha.media/en/2022/10/28/the-unique-ukrainian-situational-awareness-system-delta-was-presented-at-the-annual-nato-event/

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4, Pt.1), 377–383. https://doi.org/10.1037/h0025589

David, U., & Bode-Asa, A. (2023). *An Overview of Social Engineering : The Role of Cognitive Biases Towards Social Engineering-Based Cyber-Attacks, Impacts and Countermeasures*. https://doi.org/10.13140/RG.2.2.12421.12003

Day, P., Twiddy, J., & Dubljević, V. (2022). Present and Emerging Ethical Issues with tDCS use: A Summary and Review. *Neuroethics, 16*(1), 1. https://doi.org/10.1007/s12152-022-09508-9

De Mascarel, H. (2025). La guerre cognitive et l'intelligence économique : Une bataille pour le contrôle du réel. *Zolty Think tank.* www.zolty.fr/nos-analyses/guerre-cognitive-et-ie-controle-du-reel

De Morgny, A. (2024). Recours aux concepts et techniques de la guerre cognitive dans le champ politique français. *Ingénierie cognitique, 7*(1), 40-46.

Debidour, J., & Pelletier, P. (2024). De l'analyse d'audience au microciblage : Outil comportemental pour la guerre cognitive. *Ingénierie cognitique, 7*(1), 94-98.

DeFranco, J., DiEuliis, D., Giordano, J. (2019). Redefining Neuroweapons: Emerging Capabilities in Neuroscience and Neurotechnology. *Prism, 8.3.* https://ndupress.ndu.edu/Media/News/News-Article-View/Article/2053388/redefining-neuroweapons-emerging-capabilities-in-neuroscience-and-neurotechnolo/

Dehais, F., Tessier, C., & Chaudron, L. (2003). GHOST: experimenting conflicts countermeasures in the pilot's activity. *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 163-168.

Delaunay, J. (2024, octobre 24). La Russie, l'Iran et la Chine pourraient intensifier leurs efforts d'influence à l'approche des élections américaines, selon un nouveau rapport. *L'Observatoire de l'Europe.* www.observatoiredeleurope.com/la-russie-liran-et-la-chine-pourraient-intensifier-leurs-efforts-dinfluence-a-lapproche-des-elections-americaines-selon-un-nouveau-rapport_a47982.html

Deppe, C., & Schaal, G. S. (2024). Cognitive warfare: A conceptual analysis of the NATO ACT cognitive warfare exploratory concept. *Frontiers in Big Data, 7.* https://doi.org/10.3389/fdata.2024.1452129

Dethier, V., Heeren, A., Bouvard, M., Baeyens, C., & Philippot, P. (2017). Embracing the Structure of Metacognitive Beliefs: Validation of the French Short Version of the Metacognitions Questionnaire. *International Journal of Cognitive Therapy, 10*(3), 219-233. https://doi.org/10.1521/ijct.2017.10.3.219

Dewey, J. (2022). *How we think*

Dietrich, C. (2010). Decision Making: Factors that Influence Decision Making, Heuristics Used, and Decision Outcomes. *Inquiries Journal, 2*(02). www.inquiriesjournal.com/articles/180/decision-making-factors-that-influence-decision-making-heuristics-used-and-decision-outcomes

Donnot, J., Hauret, D., Tardan, V., & Ranc, J. (2022). Make Automation G.R.E.A.T (Again). *Proceedings of the 1st international conference on cognitive aircraft systems*, 92-95.

Dou, Y., Shu, K., Xia, C., Yu, P. S., & Sun, L. (2021). User Preference-aware Fake News Detection. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2051-2055. https://doi.org/10.1145/3404835.3462990

Downar, J., Siddiqi, S. H., Mitra, A., Williams, N., & Liston, C. (2024). Mechanisms of Action of TMS in the Treatment of Depression. Dans M. Browning, P. J. Cowen, & T. Sharp (Éds.), *Emerging Neurobiology of Antidepressant Treatments* (p. 233-277). Springer International Publishing. https://doi.org/10.1007/7854_2024_483

Drafat, N. (2023). Prise de décision dans les organisations : L'interface entre la perception du décideur et le processus normalisé par les lois en vigueurs. *Revue Française d'Economie et de Gestion, 4*(3), 386-404.

Du Cluzel, F. (2020). *Cognitive Warfare*. Innovation Hub. www.innovationhub-act.org/sites/default/files/2021-01/20210122_CW%20Final.pdf

Dubois, N. (1997). L'étude des influences psychologiques. *Les influences psychologiques - Approches scientifiques et prospectives.*

Ducourneau, A. (2024). Un design lab. Pour la sécurité cognitive. *Ingénierie cognitique, 7*(1), 88-93.

Dyèvre, A. (2015). Renseignement, facteur humain et biais cognitifs. *Revue Défense Nationale, 776*(1), 80-86. https://doi.org/10.3917/rdna.776.0080

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications, 14*(1), 62. https://doi.org/10.1038/s41467-022-35576-9

Ecole de Guerre Economique. (2001, novembre 14). *Les principes de la guerre de l'information*. www.ege.fr/infoguerre/2001/11/les-principes-de-la-guerre-de-l-information

Ehrhard, T. (2019). Big Data, algorithmes et élections. Les techniques de micro ciblage électoral sont-elles efficaces ? *XVe Congrès de l'Association Française de Science Politique, Sciences Po Bordeaux*.

www.academia.edu/43428400/Big_Data_algorithmes_et_%C3%A9lections_Les_techniques_de_micro_ciblage_%C3%A9lectoral_sont_elles_efficaces

Elbanna, S., & Child, J. (2007). The Influence of Decision, Environmental and Firm Characteristics on the Rationality of Strategic Decision-Making. *Journal of Management Studies, 44*(4), 561-591. https://doi.org/10.1111/j.1467-6486.2006.00670.x

Ellis, S., Mendel, R., & Nir, M. (2006). Learning from successful and failed experience: the moderating role of kind of after-event review. *Journal of Applied Psychology, 91*(3), 669.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors, 37*(1), 32-64. https://doi.org/10.1518/001872095779049543

Endsley, M. R. (2015). Situation Awareness Misconceptions and Misunderstandings. *Journal of Cognitive Engineering and Decision Making, 9*(1), 4-32. https://doi.org/10.1177/1555343415572631

Endsley, M., Caldwell, B. S., Chiou, E. K., Cooke, N. J., Cummings, M. L., Gonzalez, C., Lee, J. D., McNeese, N. J., Miller, C., Roth, E., Rouse, W. B., Oswald, F., Bagian, J., Burley, D., Dosher, B., Israelski, E., Lockett, J., Meshkati, N., Strickland, W., & Weinger, M. (2022). *Human-AI Teaming: State-of-the-Art and Research Needs*. National Academies of Sciences, Engineering, and Medicine - National Academies Press. https://doi.org/10.17226/26355

Endsley, M. R., & Garland, D. J. (2000). Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement, 1*(1), 3-21.

Endsley, M. R., & Kiris, E. O. (1995). The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors, 37*(2), 381 394. https://doi.org/10.1518/001872095779064555

Engin, A., & Vetschera, R. (2017). Information representation in decision making: The impact of cognitive style and depletion effects. *Decision Support Systems, 103*, 94-103. https://doi.org/10.1016/j.dss.2017.09.007

Erard, P., & Paquelet, S. (2024). La matrice DISARM, futur outil occidental contre la manipulation de l'information ? In *Lutte contre les manipulations de l'information—Regards croisés de spécialistes et d'acteurs du domaine* (Pôle d'Excellence Cyber, Vol. 2, p. 40-41).

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science, 8*(3), 223-241.

Evans-Pritchard, B. (2013). Aiming To Reduce Cleaning Costs. *Works That Work, 1.* worksthatwork.com/1/urinal-fly

Facal, J. (2011, février 16). *Les « idiots utiles ».* Le Journal de Montréal. www.journaldemontreal.com/2011/02/16/les-idiots-utiles

Fang, H., Wan, X., Zheng, S., & Meng, L. (2020). The Spillover Effect of Autonomy Frustration on Human Motivation and Its Electrophysiological Representation. *Frontiers in Human Neuroscience, 14.* https://doi.org/10.3389/fnhum.2020.00134

Fasolo, B., Heard, C., & Scopelliti, I. (2025). Mitigating Cognitive Bias to Improve Organizational Decisions : An Integrative Review, Framework, and Research Agenda. *Journal of Management, 51*(6), 2182-2211. https://doi.org/10.1177/01492063241287188

Feng, S., Wan, H., Wang, N., Li, J., & Luo, M. (2021). TwiBot-20: A Comprehensive Twitter Bot Detection Benchmark. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4485-4494. https://doi.org/10.1145/3459637.3482019

Fenstermacher, L., Uzcha, D., Larson, K., Vitiello, C., & Shellman, S. (2023). New perspectives on cognitive warfare. *Signal Processing, Sensor/Information Fusion, and Target Recognition XXXII, 12547*, 172-187. https://doi.org/10.1117/12.2666777

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fildes, J. (2010, septembre 23). Stuxnet worm « targeted high-value Iranian assets ». *BBC News.* www.bbc.com/news/technology-11388018

Fischer, K. (2009). The Influence of Neoliberals in Chile before, during, and after Pinochet. *The road from Mont Pèlerin: The making of the neoliberal thought collective*, 305-346.

Fischer, S., Itoh, M., & Inagaki, T. (2014). Identifying the cognitive causes of human error through experimentation. *Journal Européen des Systèmes Automatisés, 48*, 4-6.

Fombrun, C. J. (1996). *Reputation: Realizing Value from the Corporate Image*. Harvard Business School Press.

Fooken, J., & Schaffner, M. (2016). The Role of Psychological and Physiological Factors in Decision Making under Risk and in a Dilemma. *Frontiers in Behavioral Neuroscience, 10*, 2. https://doi.org/10.3389/fnbeh.2016.00002

Franiatte, N., Boissin, E., Delmas, A., & De Neys, W. (2023). Boosting Debiasing: Impact of Repeated Training on Reasoning. *Learning and Instruction.*

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*(4), 25-42. https://doi.org/10.1257/089533005775196732

Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology, 4*(2), 195-202. https://doi.org/10.1037/h0023552

Freedman, D., Pisani, R., & Purves, R. (2007). Statistics (international student edition). *Pisani, R. Purves, 4th Edition.*

Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. *Social Networks, 1*(3), 215-239. https://doi.org/10.1016/0378-8733(78)90021-7

French, J.R.P., Jr., & Raven, B. (1959). The bases of social power. Dans D. Cartwright (Ed.), *Studies in Social Power* (pp. 150-167). Ann Arbor, MI: Institute for Social Research.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association, 32*(200), 675-701.

Fulconis, F., & Paché, G. (2008). Le management stratégique des réseaux inter-organisationnels à l'épreuve des comportements opportunistes : Élaboration d'un cadre d'analyse. *La Revue Des Sciences de Gestion,* 35-43. https://doi.org/10.1051/LARSG:2008017

Gabrys, R. L., Tabri, N., Anisman, H., & Matheson, K. (2018). Cognitive Control and Flexibility in the Context of Stress and Depressive Symptoms: The Cognitive Control and Flexibility Questionnaire. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.02219

Gagliano, G. (2016). *Guerre économique et guerre cognitive*. Centre Français de Recherche sur le Renseignement. https://cf2r.org/tribune/guerre-economique-et-guerre-cognitive/

Gallarotti, G. M. (2011). Soft power: What it is, why it's important, and the conditions for its effective use. *Journal of Political Power, 4*(1), 25-47. https://doi.org/10.1080/2158379X.2011.557886

Galston, W. A. (2001). Political Knowledge, Political Engagement, and Civic Education. *Annual Review of Political Science, 4*(1), 217-234. https://doi.org/10.1146/annurev.polisci.4.1.217

Gamble, K. R., Vettel, J. M., Patton, D. J., Eddy, M. D., Caroline Davis, F., Garcia, J. O., Spangler, D. P., Thayer, J. F., & Brooks, J. R. (2018). Different profiles of decision making and physiology under varying levels of stress in trained military personnel. *Official Journal of the International Organization of Psychophysiology, 131*, 73-80. https://doi.org/10.1016/j.ijpsycho.2018.03.017

García-Alcaraz, J. L., Sánchez-Ramírez, C., Díaz-Reza, J. R., Avelar-Sosa, L., & Puig-i-Vidal, R. (2023). Trends on Decision Support Systems: A Bibliometric Review. Dans J. A. Zapata-Cortes, C. Sánchez-Ramírez, G. Alor-Hernández, & J. L. García-Alcaraz (Éds.), *Handbook on Decision Making* (Vol. 3, p. 169-199). Springer International Publishing. https://doi.org/10.1007/978-3-031-08246-7_8

Gass, R. H., & Seiter, J. S. (2022). *Persuasion: Social influence and compliance gaining* (Seventh edition). Routledge.

Gayard, L. (2018). Idiocratie 2.0. *Revue des Deux Mondes*, 71-76.

Gazeau-Secret, A. (2013). « Soft power » : L'influence par la langue et la culture. *Revue internationale et stratégique, 89*(1), 103-110. https://doi.org/10.3917/ris.089.0103

Genieys, W., Irondelle, B., Joana, J., Michel, L., Muller, P., Secondy, P., & Smith, A. (2003). *Groupes d'influence et processus de décision dans le domaine de la Défense—Approches comparées.* Centre d'Études Politiques de l'Europe Latine. www.academia.edu/26558433/Groupes_dinfluence_et_processus_de_d%C3%A9cision_dans_le_domaine_de_la_D%C3%A9fense

Géré, F. (2005). Mutations de la guerre psychologique. *Stratégique, N° 85*(1), 87-109. https://doi.org/10.3917/strat.085.0087

Gervais, W. M., & Norenzayan, A. (2012). Analytic Thinking Promotes Religious Disbelief. *Science, 336*(6080), 493-496. https://doi.org/10.1126/science.1215647

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making, 20*(1), 21-35.

Giordano, J. (2018, octobre 29). *Dr. James Giordano: The Brain is the Battlefield of the Future.* Modern War Institute. Consulté le 29 novembre 2022, à l'adresse www.youtube.com/watch?v=N02SK9yd60s

Girard, A. J., Reeve, C. L., & Bonaccio, S. (2016). Assessing decision-making style in French-speaking populations: Translation and validation of the general decision-making style questionnaire. *Revue Européenne de Psychologie Appliquée, 66*, 325-333.

González, F., del Val, M. P., & Cano, A. R. (2022). Systematic literature review of interpretative positions and potential sources of resistance to change in organizations. *Intangible Capital, 18*(2), 2. https://doi.org/10.3926/ic.1806

Grace, L. D., & Liang, S. (2023). Examining Misinformation and Disinformation Games Through Inoculation Theory and Transportation Theory. *Proceedings of the 56th Hawaii International Conference on System Sciences.*

Green, S. A. (2008). *Cognitive Warfare* [Joint Military Intelligence College]. https://theaugeanstables.com/wp-content/uploads/2014/04/Stuart-Green-LTC-USN-Cognitive-Warfare.pdf

Greenslade, R. (2020). Not useful, not idiots. *British Journalism Review, 31*(4), 73-75. https://doi.org/10.1177/0956474820978086b

Grint, K. (2005). Problems, problems, problems: The social construction of 'leadership'. *Human Relations, 58*(11), 1467-1494. https://doi.org/10.1177/0018726705061314

Grynszpan, E. (2024, octobre 12). The Russian disinformation machine, a constantly changing ecosystem. *Le Monde*. www.lemonde.fr/en/international/article/2024/10/12/the-russian-disinformation-machine-a-constantly-changing-ecosystem_6729193_4.html

Hammond, K. R. (2000). *Judgments Under Stress*. Oxford University Press.

Hansen, B., Flay, B. R., Phil, D., Graham, J. W., & Sobel, J. (1988). Affective and Social Influences Approaches to the Prevention of Multiple Substance Abuse among Seventh Grade Students: Results from Project SMART. *Preventive Medicine, 17*, 135-154.

Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., & Etter, M. (2011). Good Friends, Bad News—Affect and Virality in Twitter. Dans J. J. Park, L. T. Yang, & C. Lee (Éds.), *Future Information Technology* (p. 34 43). Springer. https://doi.org/10.1007/978-3-642-22309-9_5

Hanson, V. D. (2004). *Our Weird Way of War.* Hoover Institution. Consulté le 1 décembre 2022, à l'adresse www.hoover.org/research/our-weird-way-war

Harbulot, C. (2004). De la légitimité de la guerre cognitive. *Revue internationale et stratégique, 56*(4), 63-67. https://doi.org/10.3917/ris.056.0063

Harbulot, C., Moinet, N., & Lucas, D. (2002). *La guerre cognitive : A la recherche de la suprématie stratégique.* VIème forum intelligence économique de l'Association aéronautique et astronautique française, Menton.

Hardy, M. (2024). (War)gaming : Les questions pédagogiques comme enjeu de résilience cognitive. *Ingénierie cognitique, 7*(1), 81-87

Harmon-Jones, E., & Harmon-Jones, C. (2002). Testing the Action-Based Model of Cognitive Dissonance: The Effect of Action Orientation on Postdecisional Attitudes*. Personality and Social Psychology Bulletin, 28*(6), 711-723. https://doi.org/10.1177/0146167202289001

Harmon-Jones, E., & Harmon-Jones, C. (2007). Cognitive dissonance theory. Dans J. Y. Shah & W. L. Gardner (Éds.), *Handbook of Motivation Science*. The Guilford Press.

Hartley III, D. S., & Jobson, K. O. (2020). *Cognitive Superiority: Information to power, the road to winning in the Sixth Domain*. Hartley Consulting.

Hartono, J. (2012). The recency effect of accounting information. *Gadjah Mada International Journal of Business, 6*(1), 1. https://doi.org/10.22146/gamaijb.5536

Haushofer, J., Fehr, E. (2014). On the psychology of poverty. *Science, 344*(6186), 862-867. https://doi.org/10.1126/science.1232491

Heath, R. L., & Palenchar, M. J. (2009). *Strategic Issues Management: Organizations and Public Policy Challenges* (2e éd.). SAGE Publications. sk.sagepub.com/book/mono/strategic-issues-management-2e/toc

Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons Inc. https://doi.org/10.1037/10628-000

Heilbrunn, B. (2022). Manipulation, influence, consentement : de quoi parle-t-on ? *Les Grands Dossiers des Sciences Humaines, N° 66*(3), 6-9. https://doi.org/10.3917/gdsh.066.0006

Henderson, J. C., & Nutt, P. C. (1980). The Influence of Decision Style on Decision Making Behavior. *Management Science, 26*(4), 371-386. https://doi.org/10.1287/mnsc.26.4.371

HFM Exploratory Team 356. (2023). *Mitigating and Responding to Cognitive Warfare.* NATO STO.

Higashijima, M., Kadoya, H., & Yanai, Y. (2024). The Dynamics of Electoral Manipulation and Institutional Trust in Democracies: Election Timing, Blatant Fraud, and the Legitimacy of Governance. *Public Opinion Quarterly, 88*, 472-494.

Higgins, E. T. (2011). Regulatory focus theory. Dans P. A. M. Van Lange, E. T. Higgins, & A. W. Kruglanski, *Handbook of Theories of Social Psychology: Volume 1* (p. 483-504). Sage Publications. https://www.torrossa.com/en/resources/an/5017496

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin, 138*(2), 211-237. https://doi.org/10.1037/a0025940

Holeindre, J.-V. (2017). *La ruse et la force : Une autre histoire de la stratégie*. Perrin.

Hollnagel, E. (1993). *Human Reliability Analysis: Context and Control.* Academic Press.

Hristakieva, K., Cresci, S., Martino, G. D. S., Conti, M., & Nakov, P. (2022). The Spread of Propaganda by Coordinated Communities on Social Media. *14th ACM Web Science Conference 2022*, 191-201. https://doi.org/10.1145/3501247.3531543

Hristova, D., Musolesi, M., & Mascolo, C. (2014). Keep Your Friends Close and Your Facebook Friends Closer: A Multiplex Network Approach to the Analysis of Offline and Online Social Ties. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 206-215.

Hsu, C.-T., Sims, T., & Chakrabarti, B. (2018). How mimicry influences the neural correlates of reward: An fMRI study. *Neuropsychologia, 116*, 61-67. https://doi.org/10.1016/j.neuropsychologia.2017.08.018

Huang G., & Wang S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. *Journal of Communication, 73*(6), 552–562. https://doi.org/10.1093/joc/jqad024

Huber, O., Huber, O. W., & Bär, A. S. (2011). Information search and mental representation in risky decision making: The advantages first principle. *Journal of Behavioral Decision Making, 24*(3), 223-248. https://doi.org/10.1002/bdm.674

Hübert, R., & Little, A. T. (2020). *Kompromat.*

Hui, C. H. (1988). Measurement of individualism-collectivism. *Journal of Research in Personality, 22*(1), 17-36. https://doi.org/10.1016/0092-6566(88)90022-0

Hung, T.-C., & Hung, T.-W. (2022). How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars. *Journal of Global Security Studies, 7*(4). https://doi.org/10.1093/jogss/ogac016

Inkster, N. (2016). Information Warfare and the US Presidential Election. *Survival, 58*(5), 23-32. https://doi.org/10.1080/00396338.2016.1231527

Isaac, H., Campoy, E., & Kalika, M. (2007). Surcharge informationnelle, urgence et TIC. L'effet temporel des technologies de l'information. *Management & Avenir, 13*(3), 149-168. https://doi.org/10.3917/mav.013.0149

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer, 51*(8), 56-59. https://doi.org/10.1109/MC.2018.3191268

Jaesa. (2025, avril 24). Palantir : Le Maven Smart System (MSS NATO). *Intelligence Artificielle et Transhumanisme.* https://iatranshumanisme.com/2025/04/24/palantir-le-maven-smart-system-mss-nato/

Janin, P. (2024). Influence des réseaux sociaux sur la résilience cognitive des jeunes – impact sur les combattants. *Ingénierie cognitique, 7*(1), 99-108.

Jin, H., Hou, L.-J., Wang, Z.-G. (2018). Military Brain Science – How to influence future wars. *Chinese Journal of Traumatology*, *21*(5), 277-280. https://doi.org/10.1016/j.cjtee.2018.01.006

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big Five Inventory. *Journal of personality and social psychology*.

Johns Hopkins University & Imperial College London. (2021). *L'OTAN et la biotechnologie cognitive : Questions et perspectives*. NATO Review. www.nato.int/docu/review/fr/articles/2021/02/26/lotan-et-la-biotechnologie-cognitive-questions-et-perspectives/index.html

Jolicard, A.-E., & Gardin, A. (2024). Anticiper le risque de manipulation de l'information : Le scoring de risque selon l'exposition d'information personnelle. Dans *Lutte contre les manipulations de l'information—Regards croisés de spécialistes et d'acteurs du domaine* (Pôle d'Excellence Cyber, Vol. 2). https://www.pole-excellence-cyber.org/evenements/lutte-contre-les-manipulations-de-linformation-decouvrez-le-tome-2/

Joule, R.-V., & Beauvois, J.-L. (2010). *La soumission librement consentie. Comment amener les gens à faire librement ce qu'ils doivent faire ?* Presses universitaires de France. https://shs.cairn.info/la-soumission-librement-consentie--9782130578826

Joule, R. V., Beauvois, J. L. (1987). *Petit traité de manipulation à l'usage des honnêtes gens.* Presses universitaires de Grenoble.

Jowett, G. S., & O'Donnell, V. (1986). *Propaganda & Persuasion.* SAGE Publications.

Kahneman, D., Lovallo, D., & Sibony, O. (2011, juin). The Big Idea: Before You Make That Big Decision…. *Harvard Business Review*. https://hbr.org/2011/06/the-big-idea-before-you-make-that-big-decision

Kammers, M. P. M., de Vignemont, F., Verhagen, L., & Dijkerman, H. C. (2009). The rubber hand illusion in action. *Neuropsychologia, 47*(1), 204-211. https://doi.org/10.1016/j.neuropsychologia.2008.07.028

Keen, P. G. W., & Scott Morton, M. S. (1978). *Decision support systems: An organizational perspective.* Addison-Wesley Pub. Co. https://cir.nii.ac.jp/crid/1970586434902171824

Keshavarz, H. (2020). Evaluating credibility of social media information: Current challenges, research directions and practical criteria*. Information Discovery and Delivery, 49*(4), 269 279. https://doi.org/10.1108/IDD-03-2020-0033

Ketelaars, E., Gaudin, C., Flandin, S., & Poizat, G. (2024). Resilience training for critical situation management. An umbrella and a systematic literature review. *Safety Science, 170*, 106311. https://doi.org/10.1016/j.ssci.2023.106311

Khaled, S., El-Tazi, N., & Mokhtar, H. M. O. (2018). Detecting Fake Accounts on Social Media. *2018 IEEE International Conference on Big Data*, 3672-3681. https://doi.org/10.1109/BigData.2018.8621913

Kirchner, J., & Reuter, C. (2020). Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), article 140: 1-27. https://doi.org/10.1145/3415211

Kodalle, T., & Ormrod, D. D. (2023). *A General Theory of Influence in a DIME/PMESII/ASCOP/IRC² Model.*

Kojima, M. (2024). Organizational or Individual? The Effect of Social Networks on Volunteer Activities in Japan. *International Journal of Voluntary and Nonprofit Organizations, 35*(3), 583-596. https://doi.org/10.1007/s11266-023-00623-6

Korteling, J. E. (Hans), Paradies, G. L., & Sassen-van Meer, J. P. (2023). Cognitive bias and how to improve sustainable decision making. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1129835

Kosinski, M. (2019). *Computational Psychology*. Advanced social psychology: The state of the science.

Košková, Z. (2025). *TikTok as a Tool of Cognitive Warfare?* Adapt Institute. www.adaptinstitute.org/wp-content/uploads/2025/03/Zuzana-Koskova-Tiktok.pdf

Kozloski, R. P. (2018, février 1). Knowing Yourself Is Key in Cognitive Warfare. *RealClearDefense*. Consulté le 16 décembre 2022, à l'adresse www.realcleardefense.com/articles/2018/02/01/knowing_yourself_is_key_in_cognitive_warfare_112992-full.html

Krishnan, A. (2017). *Military Neuroscience and the Coming Age of Neurowarfare*. Taylor & Francis.

Kumar, A. (2024). Chinese Cognitive Warfare: Exploring a Framework and Identifying Contours Abhishek Kumar. *Defence & Diplomacy, 13*(3), 39-53.

Kumar, N., & Dixit, A. (2019). Role of Nanotechnology in Futuristic Warfare. Dans *Nanotechnology for Defence Applications* (p. 301-329). Springer International Publishing. https://doi.org/10.1007/978-3-030-29880-7_8

Kuutila, M., Kiili, C., Kupiainen, R., Huusko, E., Li, J., Hosio, S., Mäntylä, M., Coiro, J., & Kiili, K. (2024). Revealing complexities when adult readers engage in the credibility evaluation of social media posts. *Computers in Human Behavior*, 151. https://doi.org/10.1016/j.chb.2023.108017

Lagadec, P. (2010). La force de réflexion rapide — Aide au pilotage des crises. *Préventique Sécurité*, 31-35.

Larrick, R. P. (2004). Debiasing. Dans D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–337). Blackwell Publishing. https://doi.org/10.1002/9780470752937.ch16

Lau, J. M. H., Rashid, A. J., Jacob, A. D., Frankland, P. W., Schacter, D. L., & Josselyn, S. A. (2020). The role of neuronal excitability, allocation to an engram and memory linking in the behavioral generation of a false memory in mice. N*eurobiology of Learning and Memory, 174*, 107284. https://doi.org/10.1016/j.nlm.2020.107284

Laudy, C., Mattioli, J., & Museux, N. (2006). *Cognitive Situation Awareness for Information Superiority*. IST Panel on Information Fusion for Command Support. www.researchgate.net/publication/235046135_Cognitive_Situation_Awareness_for_Information_Superiority

Lazega, E. (1994). Analyse de réseaux et sociologie des organisations. *Revue française de sociologie, 35*(2), 293-320. https://doi.org/10.2307/3322036

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

Le Blanc, B. (2018). Cognition et cognitique. *Hermès, La Revue, 80*(1), 37-38. https://doi.org/10.3917/herm.080.0037

Le Guyader, H. (2020). *Weaponization of neurosciences.* École Nationale Supérieure de Cognitique, Bordeaux, France. innovationhub-act.org/wp-content/uploads/2023/12/WoNS.pdf

Le Guyader, H., Cole, A. (2020). *Cognitif, un Sixième Domaine d'Opérations ? FICINT document*. Norfolk VA, USA : NATO ACT Innovation Hub. www.innovationhub-act.org/sites/default/files/2021-04/FR%20version%20v6.pdf

Le Monde avec AFP. (2023, février 7). Réforme des retraites : Des députées RN cibles de messages d'intimidation. *Le Monde.* Consulté le 25 octobre 2023, à l'adresse www.lemonde.fr/politique/article/2023/02/07/reforme-des-retraites-des-deputees-rn-cibles-de-messages-d-intimidation_6160845_823448.html

Lebraty, J.-F. (2004). Biais cognitifs : Quel statut dans la prise de décision assistée ? *Systèmes d'information et management, 9*(3), 87-116.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors, 46*(1), 50-80.

Lehto, M. (2022). Cyber-Attacks Against Critical Infrastructure. Dans M. Lehto & P. Neittaanmäki (Éds.), *Cyber Security: Critical Infrastructure Protection* (p. 3-42). Springer International Publishing. https://doi.org/10.1007/978-3-030-91293-2_1

Leino, T. (2022). Informal Leadership: An Integrative View and Future Research. In: H. Katajamäki, M. Enell-Nilsson, H. Kauppinen-Räisänen & H. Limatius (Eds.). *Responsible Communication*. VAKKI Publications 14. 118–136. ISBN 978-952-69732-1-0.

Leloup, D., Reynaud, F. (2023). Milliardaires, lanceurs d'alerte, criminels, opposants politiques : Les cibles de l'officine de désinformation « Team Jorge ». *Le Monde.* Consulté le 15 février 2023, à l'adresse www.lemonde.fr/pixels/article/2023/02/15/milliardaires-lanceurs-d-alerte-criminels-opposants-politiques-les-cibles-d-une-usine-a-fake-news_6161841_4408996.html

Lemaire, P. (1999). *Psychologie cognitive.* De Boeck Université.

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition, 6*(4), 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008

Li, R., & Suh, A. (2015). Factors Influencing Information credibility on Social Media Platforms: Evidence from Facebook Pages. *Procedia Computer Science, 72*, 314-328. https://doi.org/10.1016/j.procs.2015.12.146

Limonier, K., & Audinet, M. (2017). La stratégie d'influence informationnelle et numérique de la Russie en Europe *Hérodote, N° 164*(1), 123-144. https://doi.org/10.3917/her.164.0123

List, J. A., Ramirez, L. M., Seither, J., Unda, J., & Vallejo, B. H. (2024). Critical thinking and misinformation vulnerability: Experimental evidence from Colombia. *PNAS Nexus, 3*(10). https://doi.org/10.1093/pnasnexus/pgae361

Loewenstein, G., & Chater, N. (2017). Putting nudges in perspective. *Behavioural Public Policy, 1*(1), 26-53. https://doi.org/10.1017/bpp.2016.7

Lopez, J., Perumalla, K., & Siraj, A. (2021). The Overton Window: A Tool for Information Warfare. Dans *ICCWS 2021 16th International Conference on Cyber Warfare and Security* (p. 20-27). Academic Conferences Limited.

Lorenc, E. S., Mallett, R., & Lewis-Peacock, J. A. (2021). Distraction in Visual Working Memory: Resistance is Not Futile. *Trends in Cognitive Sciences, 25*(3), 228-239. https://doi.org/10.1016/j.tics.2020.12.004

Madsen, S. R., John, C. R., & Miller, D. (2006). Influential Factors in Individual Readiness for Change. *Journal of Business and Management, 12*(2), 93-110. https://doi.org/10.1504/JBM.2006.141142

Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., He, J., & van der Linden, S. (2023). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods.* https://doi.org/10.3758/s13428-023-02124-2

Malin, I. (2022). *Cyberattaques : comment l'Ukraine a failli perdre la guerre avant même l'invasion russe*. France TV Info. Consulté le 5 octobre 2022, à l'adresse www.francetvinfo.fr/internet/securite-sur-internet/cyberattaques/video-cyberattaques-comment-l-ukraine-a-failli-perdre-la-guerre-avant-meme-l-invasion-russe_5397346.html#xtor=CS2-765-%5Bautres%5D-

Margraff, G. (2024). Réseaux sociaux et guerre cognitive : La menace des algorithmes de recommandation de contenus. *Sécurité et stratégie, 37*(2), 62-66. https://doi.org/10.3917/sestr.037.0062

Marjanović, A., & Smiljanić, D. (2025). Cognitive Warfare—The human mind as the new battlefield. *Proceedings of the Defence and Security Conference, 1*(1), 88-114.

Marpaung, A. Y. (2025, avril 27). Made in China, but Not Free in China: Why is TikTok Restricted? *Modern Diplomacy*. moderndiplomacy.eu/2025/04/27/made-in-china-but-not-free-in-china-why-is-tiktok-restricted

Marsili, M. (2021). The Russian Influence Strategy in Its Contested Neighbourhood. Dans H. Mölder, V. Sazonov, A. Chochia, & T. Kerikmäe (Éds.), *The Russian Federation in Global Knowledge Warfare: Influence Operations in Europe and Its Neighbourhood* (p. 149-172). Springer International Publishing. https://doi.org/10.1007/978-3-030-73955-3_8

Martínez, N., Agudo, U., & Matute, H. (2022). Human cognitive biases present in Artificial Intelligence. *Revista Internacional de Los Estudios Vascos, 67*(2). https://www.eusko-ikaskuntza.eus/zbk/TAG_Nzenbaki_VALUE/

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading, 12*(8), 639-646.

McCreight, R. (2022). Neuro-Cognitive Warfare: Inflicting Strategic Impact via Non-Kinetic Threat. *Small Wars Journal*. https://smallwarsjournal.com/jrnl/art/neuro-cognitive-warfare-inflicting-strategic-impact-non-kinetic-threat

McCreight, R. (2024). The war inside your mind: Unprotected brain battlefields and neuro-vulnerability. *Academia Biology, 2*(1). www.academia.edu/2837-4010/2/1/10.20935/AcadBiol6156

McDougall, J., Zezulkova, M., van Driel, B., & Sternadel, D. (2018). *Teaching media literacy in Europe: Evidence of effective school practices in primary and secondary education* (NESET II report). Luxembourg: Publications Office of the European Union. eprints.bournemouth.ac.uk/31574/1/AR2_Teaching%20Media%20Literacy_NESET.pdf

McGovern, A. (2021, mai 27). *Artificial intelligence system could help counter the spread of disinformation*. MIT News. news.mit.edu/2021/artificial-intelligence-system-could-help-counter-spread-disinformation-0527

McIntyre, L. (2023). *On Disinformation: How to Fight for Truth and Protect Democracy*. MIT Press.

Mercier, H. (2017). Confirmation bias - Myside bias. Dans *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory, 2nd ed* (p. 99-114). Routledge/Taylor & Francis Group.

Metz, M. (2021). Overview of Change in Organizations. Resistance to Change. A Literature Review. *« Ovidius » University Annals, Economic Sciences Series, XXI*(1).

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & Mccann, R. M. (2003). Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment.

Annals of the International Communication Association, 27(1), 293 335. https://doi.org/10.1080/23808985.2003.11679029

Michalak, S. (2011). Motives of espionage against ones own country in the light of idiographic studies. *Polish Psychological Bulletin, 42*(1), 1-4. https://doi.org/10.2478/v10059-011-0001-2

Micheletti, M. (2003). *Political Virtue and Shopping: Individuals, Consumerism, and Collective Action*. Palgrave Macmillan US. https://doi.org/10.1057/9781403973764

Milgram, S. (1963). Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology, 67*(4), 371-378. https://doi.org/10.1037/h0040525

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How Can Decision Making Be Improved? *Perspectives on Psychological Science, 4*(4), 379-383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Miller, S. (2023). Cognitive warfare: An ethical analysis. *Ethics and Information Technology, 25*(3), 46. https://doi.org/10.1007/s10676-023-09717-7

Milstein, M. (2020). The Cognitive Campaign: Myth vs. Reality. *Strategic Assessment, 23*(2).

Mitnick, K. D., & Simon, W. L. (2002). *The Art of Deception: Controlling the human element of security*. John Wiley & Sons.

Mohlin, M. (2014). Commercialisation of Warfare and Shadow Wars: Private Military Companies as Strategic Tools. *St Antony's International Review, 9*(2), 24-38.

Montocchio, P. (2021). Avant-propos par le directeur adjoint du Collaboration Support Office (CSO) STO. Dans B. Claverie, B. Prébot et F. du Cluzel (dir.), *La Guerre Cognitique* (p. vii-viii). CSO

Moran, N. (2020). Illusion of safety: How consumers underestimate manipulation and deception in online (vs. offline) shopping contexts. *Journal of Consumer Affairs, 54*(3), 890-911. https://doi.org/10.1111/joca.12313

Morelle-Gerritsen, M., Marion, D., Cegarra, J., Unrein, H., Letouzé, T., & André, J.-M. (2025). Evaluating and Influencing Strategy in Real-time: Example of a Collaborative Strategy Game. Dans I. Wiafe, A. Babiker, J. Ham, K. Oyibo, & E. Vlahu-Gjorgievska (Éds.), *Persuasive Technology. PERSUASIVE 2025 Satellite Events.* Springer Nature Link. https://link.springer.com/book/9783031971761

Moreno, J. D. (2012). *Mind Wars: Brain Science and the Military in the 21st Century*. Bellevue Literary Press.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions: Improved Decision Making With a Single Training Intervention. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 129-140. https://doi.org/10.1177/2372732215600886

Morgan, D. L. (1997). *Focus Groups as Qualitative Research*. SAGE.

Morrison, J. E., & Fletcher, J. D. (2002). Cognitive Readiness. *Institute for Defense Analyses.*

Mueller, R. S. (2019). *Report on the Investigation into Russian Interference in the 2016 Presidential Election* (Washington, DC: US Department of Justice, Vol. 1).

Muhammad, Z. (2024, décembre 18). Introduction à la guerre cognitive : Les enjeux d'une nouvelle forme de conflictualité. *CEDIRE*. www.cedire.fr/articles/introduction-a-la-guerre-cognitive-les-enjeux-dune-nouvelle-forme-de-conflictualite

Mwai, P. (2022, mai 3). Charnier de Gossi : Quelles sont les accusations de la France concernant le charnier découvert au Mali ? *BBC News Afrique*. www.bbc.com/afrique/monde-61307075

Nalepa, M., & Sonin, K. (2020). How Does Kompromat Affect Politics ? A Model of Transparency Regimes. *CEPR Discussion Paper No. DP14992.*

Nasi, M. (2023). A l'Ecole de guerre économique, une formation pour agents secrets de la mondialisation. *Le Monde.* Consulté le 13 avril 2023, à l'adresse www.lemonde.fr/campus/article/2023/04/05/a-l-ecole-de-guerre-economique-une-formation-pour-agents-secrets-de-la-mondialisation_6168303_4401467.html

NATO. (s. d.). Media—(Dis)information—Security: Information Warfare. Consulté le 24 janvier 2023, à l'adresse https://www.nato.int/nato_static_fl2014/assets/pdf/2020/5/pdf/2005-deepportal4-information-warfare.pdf

Nechaev, V., & Kosyakov, S. (2024). Non-autoregressive real-time Accent Conversion model with voice cloning. *Pre-print - Ivanovo State Power Engineering University.*

Nemenyi, P. B. (1963). *Distribution-free multiple comparisons.* Princeton University.

Nemeth, C. J. (1995). Dissent as driving cognition, attitudes, and judgments. *Social Cognition, 13*(3), 273-291. https://doi.org/10.1521/soco.1995.13.3.273

Newman, E. J., & Schwarz, N. (2024). Misinformed by images: How images influence perceptions of truth and what can be done about it. *Current Opinion in Psychology, 56*, 101778. https://doi.org/10.1016/j.copsyc.2023.101778

Newman, N., Fletcher, R., Robertson, C. T., Arguedas, A. R., & Nielsen, R. K. (2024). *Reuters Institute Digital News Report 2024.*

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259. https://doi.org/10.1037/0033-295X.84.3.231

Norris, P. (2014). *Why Electoral Integrity Matters*. Cambridge University Press.

Nye, J. S. (1990). Soft Power. *Foreign Policy, 80*, 153-171. https://doi.org/10.2307/1148580

Oates, K., & Wilson, M. (2002). Nominal kinship cues facilitate altruism. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 269*(1487), 105-109. https://doi.org/10.1098/rspb.2001.1875

Ofcom. (2025). *Misinformation and Disinformation: Literature Review.* www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/mis-and-disinformation/misinformation-and-disinformation-literature-review.pdf?v=397787

Office of the Director of National Intelligence. (2022). *Complementary Efforts on Anomalous Health Incidents*. www.dni.gov/index.php/newsroom/reports-publications/reports-publications-2022/item/2273-complementary-efforts-on-anomalous-health-incidents

Oliver, C. (1991). Strategic Responses to Institutional Processes. *The Academy of Management Review, 16*(1), 145-179. https://doi.org/10.2307/258610

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group.

Onwujekwe, G., & Weistroffer, H. R. (2025). Intelligent Decision Support Systems: An Analysis of the Literature and a Framework for Development. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-024-10571-1

Orasanu, J., Calderwood, R., & Zsambok, C. E. (1993). *Decision making in action: Models and methods* (Vol. 3). G. A. Klein (Ed.). Norwood, NJ: Ablex.

Oreg, S. (2003). Resistance to change: Developing an individual differences measure. *Journal of Applied Psychology, 88*(4), 680-693. https://doi.org/10.1037/0021-9010.88.4.680

Oreskes, N., & Conway, E. M. (2010). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press.

Orinx, K., & Struye de Swielande, T. (2021). Chapitre 8 – La guerre cognitive – Pourquoi l'Occident pourrait perdre face à la Chine ? Dans B. Claverie, B. Prébot et F. du Cluzel (dir.), *La Guerre Cognitique* (p. 8.1-8.7). CSO

Orr, G. (2003). *Diffusion of innovations, by Everett Rogers (1995).*

Orr, G. E. (1983). *Combat Operations C3I: Fundamentals and Interactions.* Airpower Research Institute, Air University Press.

Osoba, O. A., & Welser, W. (2017). *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence.* Rand Corporation.

Pagès, J. (2014). *Multiple factor analysis by example using R.* CRC Press.

Patino, B. (2023, novembre 27). *S'informer à l'heure des réseaux sociaux*. Vie publique – République Française. http://www.vie-publique.fr/parole-dexpert/291804-sinformer-lheure-des-reseaux-sociaux-par-bruno-patino

Pescetelli, N., Hauperich, A. K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition, 215*, 104810.

Pescosolido, A. T. (2001). Informal Leaders and the Development of Group Efficacy. *Small Group Research, 32*(1), 74-93. https://doi.org/10.1177/104649640103200104

Petkus, D. A. (2010). Ethics of human intelligence operations: Of mice and men. *International Journal of Intelligence Ethics, 1*(1), 97-121.

Petty, R. E., & Briñol, P. (2011). The Elaboration Likelihood Model. Dans P. Van Lange, A. Kruglanski, & E. Higgins, *Handbook of Theories of Social Psychology: Volume 1* (p. 224-245). SAGE Publications Ltd. https://doi.org/10.4135/9781446249215.n12

Petty, R. E., & Cacioppo, J. T. (1984). Source factors and the elaboration likelihood model of persuasion. *Advances in Consumer Research, 11*, 668-672.

Pfister, H.-R., & Böhm, G. (2008). The multiplicity of emotions: A framework of emotional functions in decision making. *Judgment and Decision Making, 3*(1), 5-17. https://doi.org/10.1017/S1930297500000127

Pham, V. K., Pham Thi, T. D., & Duong, N. T. (2024). A Study on Information Search Behavior Using AI-Powered Engines : Evidence From Chatbots on Online Shopping Platforms. *Sage Open, 14*(4). https://doi.org/10.1177/21582440241300007

Piksa, M., Noworyta, K., Piasecki, J., Gwiazdzinski, P., Gundersen, A. B., Kunst, J., & Rygula, R. (2022). Cognitive Processes and Personality Traits Underlying Four Phenotypes of Susceptibility to (Mis)Information. *Frontiers in Psychiatry, 13*. https://doi.org/10.3389/fpsyt.2022.912397

Pinard Legry, O. (2022). Neurosciences et sciences cognitives : Comment se préparer à la guerre des cerveaux ? *Revue Défense Nationale*, *N° Hors-série*(HS3), 58-76. https://doi.org/10.3917/rdna.hs09.0058

Plaisant, O., Courtois, R., Réveillère, C., Mendelsohn, G. A., & John, O. P. (2010). Validation par analyse factorielle du Big Five Inventory français (BFI-Fr). Analyse convergente avec le NEO-PI-R. *Annales Médico-psychologiques, revue psychiatrique, 168*(2), 97-106. https://doi.org/10.1016/j.amp.2009.09.003

Porter, P. (2006). Review Essay: Shadow Wars: Asymmetric Warfare in the Past and Future. *Security Dialogue, 37*(4), 551-561. https://doi.org/10.1177/0967010606072498

Prébot, B. (2020). *Représentation partagée et travail collaboratif en contexte C2 : Monitoring d'opérateurs en situation simulée de command and control*. [Thèse de doctorat]. Université de Bordeaux.

Prichard, E. C. (2021). Is the Use of Personality Based Psychometrics by Cambridge Analytical Psychological Science's "Nuclear Bomb" Moment? *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.581448

Priebe, M., O'Mahony, A., Frederick, B., Demus, A., Lin, B., Grisé, M., Eaton, D., & Doll, A. (2021). *Operational Unpredictability and Deterrence: Evaluating Options for Complicating Adversary Decisionmaking.* https://www.rand.org/pubs/research_reports/RRA448-1.html

Prier, J. (2020). Commanding the trend: Social media as information warfare. Dans *Information Warfare in the Age of Cyber Conflict* (p. 88-113). Routledge.

Qureshi, L. Z. (2008). *Nixon, Kissinger, and Allende: US involvement in the 1973 coup in Chile*. Rowman & Littlefield.

Rabat, A., Cutsem, J. V., Marcora, S. M., Lambert, A., Markwald, R., Kubala, A. G., & Friedl, K. E. (2025). Fatigue and management of warfighter mental endurance. *BMJ Military Health*. https://doi.org/10.1136/military-2025-002963

Rademacher, L., Kraft, D., Eckart, C., & Fiebach, C. J. (2023). Individual differences in resilience to stress are associated with affective flexibility. *Psychological Research, 87*(6), 1862-1879. https://doi.org/10.1007/s00426-022-01779-4

Rand Corporation. (2018). *Returning To Human Fundamentals of War*. Report for the Army. https://www.rand.org/content/dam/rand/pubs/research_briefs/RB10000/RB10040/RAND_RB10040.pdf

Raman, G., AlShebli, B., Waniek, M., Rahwan, T., & Peng, J. C. H. (2020). How weaponizing disinformation can bring down a city's power grid. *Plos one, 15*(8), e0236517.

Rasmussen, J. (1986). *Information Processing and Human-Machine Interaction. An Approach to Cognitive Engineering* (North-Holland)

Rastogi, S., & Bansal, D. (2023). A review on fake news detection 3T's: Typology, time of detection, taxonomies. *International Journal of Information Security, 22*(1), 177 212. https://doi.org/10.1007/s10207-022-00625-3

Rathi, S., Verma, J. P., Jain, R., Nayyar, A., & Thakur, N. (2022). Psychometric profiling of individuals using Twitter profiles: A psychological Natural Language Processing based approach. *Concurrency and Computation: Practice and Experience, 34*(19), e7029. https://doi.org/10.1002/cpe.7029

Ravailhe, N. (2024). Guerre cognitive et stratégies d'influence dans l'Union européenne. *Ingénierie cognitique, 7*(1), 32-39.

Rey, A., Le Goff, K., Abadie, M., & Courrieu, P. (2020). The primacy order effect in complex decision making. *Psychological Research, 84*(6), 1739-1748. https://doi.org/10.1007/s00426-019-01178-2

Rid, T. (2020). *Active Measures: The Secret History of Disinformation and Political Warfare.*

Riggio, R. E., & Newstead, T. (2023). Crisis Leadership. *Annual Review of Organizational Psychology and Organizational Behavior, 10*(Volume 10, 2023), 201-224. https://doi.org/10.1146/annurev-orgpsych-120920-044838

Rivière, A. L. (2017). Confiance et pratiques informationnelles des chefs militaires. *Revue COSSI, 2*(2). https://doi.org/10.34745/numerev_1584

Roberts, M. E. (2018). *Censored: Distraction and diversion inside China's great firewall*. Princeton University Press.

Robin, W. J. (2022). *TikTok est-il une menace pour la jeunesse ?* École de Guerre Économique. Consulté le 16 décembre 2022, à l'adresse www.ege.fr/infoguerre/tiktok-est-il-une-menace-pour-la-jeunesse

Robinson, M., Jones, K., & Janicke, H. (2015). Cyber warfare: Issues and challenges. *Computers & Security, 49*, 70-94. https://doi.org/10.1016/j.cose.2014.11.007

Roco, M. C., & Bainbridge, W. S. (Éds.). (2003). *Converging Technologies for Improving Human Performance*. Springer Netherlands. https://doi.org/10.1007/978-94-017-0359-8

Rogers, E. M. (1995). *Diffusion of Innovations* (4e éd.). Free Press.

Rohrlich, P. E. (1987). Economic culture and foreign policy: The cognitive analysis of economic policy making. *International Organization, 41*(1), 61-92. https://doi.org/10.1017/S0020818300000746

Rosen, G., Harbath, K., Gleicher, N., Leathern, R. (2019, octobre 21). Helping to Protect the 2020 US Elections. *Meta.* Consulté le 16 février 2023, à l'adresse https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/

Ross, L. (1977). The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process. *Advances in Experimental Social Psychology, 10*, 173-220. https://doi.org/10.1016/S0065-2601(08)60357-3

Rothstein, B. (2013). Corruption and Social Trust: Why the Fish Rots from the Head Down. *Social Research: An International Quarterly, 80*(4), 1009-1032.

Rugg, G., & McGeorge, P. (1997). The sorting techniques: A tutorial paper on card sorts, picture sorts and item sorts. Expert Systems, 14(2), 80-93. https://doi.org/10.1111/1468-0394.00045

Russell, N. J. (2006). *An Introduction to the Overton Window of Political Possibilities*. Mackinac Center. Consulté 1 juin 2025, à l'adresse www.mackinac.org/7504

Saade, T. (2025, janvier 30). *Election Interference in An Age of AI-Enabled Cyberattacks and Information*. Stanford International Policy Review. fsi.stanford.edu/sipr/content/election-interference-age-ai-enabled-cyberattacks-and-information-manipulation-campaigns

Salahdine, F., & Kaabouch, N. (2019). Social Engineering Attacks: A Survey. *Future Internet, 11*(4), 4. https://doi.org/10.3390/fi11040089

Salamanos, N., Jensen, M. J., He, X., Chen, Y., & Sirivianos, M. (2019). On the influence of twitter trolls during the 2016 US Presidential Election. *preprint arXiv:1910.00531.*

Santini, R. M., Agostini, L., Barros, C. E., Carvalho, D., de Rezende, R. C., Salles, D. G., Seto, K., Terra, C., & Tucci, G. (2018). Software power as soft power. *Partecipazione e Conflitto - The Open Journal of Sociopolitical Studies*. https://doi.org/10.1285/i20356609v11i2p332

Schraw, G., & Dennison, R. S. (1994). Assessing Metacognitive Awareness. *Contemporary Educational Psychology, 19*(4), 460-475. https://doi.org/10.1006/ceps.1994.1033

Scott, S. G., & Bruce, R. A. (1995). Decision-Making Style: The Development and Assessment of a New Measure. *Educational and Psychological Measurement, 55*(5), 818-831. https://doi.org/10.1177/0013164495055005017

Sedogbo, C., Mattioli, J., Laudy, C., & Museux, N. (2007). *Information fusion: A key issue for Cognitive Decision Support.* THALES Research & Technology Advance Software Department.

Segura, L. (2018). *La terrible Ventana de Overton (como legalizar cualquier cosa).* https://adelantelafe.com/la-terrible-ventana-overton-legalizar-cualquier-cosa

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin, 134*(2), 207-222. https://doi.org/10.1037/0033-2909.134.2.207

Shaima, M., Nabi, N., Rana, N. U., Islam, T., Ahmed, E., Islam, M., Mukti, M. H., & Saad-Ul-Mosaher, Q. (2024). Elon Musk's Neuralink Brain Chip: A Review on 'Brain-Reading' Device. *Journal of Computer Science and Technology Studies, 6*(1), 200-203. https://doi.org/10.32996/jcsts

Shaji, R. S., Sachin Dev, V., Brindha, T. (2019). A methodological review on attack and defense strategies in cyber warfare. *Wireless Networks, 25*(6), 3323-3334. https://doi.org/10.1007/s11276-018-1724-1

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3-4), 591-611.

Shariff, S. M. (2020). A Review on Credibility Perception of Online Information. *14th International Conference on Ubiquitous Information Management and Communication*, 1 7. https://doi.org/10.1109/IMCOM48794.2020.9001724

Shtepa, V. (2021). *Advisor to Russian Defense Minister Warns of 'Mental War': Who Is Waging It and Against Whom?* RealClear Defense. Consulté le 25 novembre 2022, à l'adresse www.realcleardefense.com/articles/2021/04/17/advisor_to_russian_defense_minister_warns_of_mental_war_who_is_waging_it_and_against_whom_773187.html

Simion, O., & Dorard, G. (2020). L'usage problématique des réseaux sociaux chez les jeunes adultes : Quels liens avec l'exposition de soi, l'estime de soi sociale et la personnalité ? *Psychologie Française, 65*(3), 243-259. https://doi.org/10.1016/j.psfr.2019.05.001

Simon, H. A. (1979). Rational Decision Making in Business Organizations. *The American Economic Review, 69*(4), 493-513.

Slater, M. D., & Rouner, D. (1996). How message evaluation and source attributes may influence credibility assessment and belief change. *Journalism & Mass Communication Quarterly.* https://doi.org/10.1177/107769909607300415

Slater, M., Spanlang, B., Sanchez-Vives, M. V., & Blanke, O. (2010). First Person Experience of Body Transfer in Virtual Reality. *PLOS ONE, 5*(5). https://doi.org/10.1371/journal.pone.0010564

Smith, P. B. (2023). Response Bias(es). Dans F. Maggino (Éd.), *Encyclopedia of Quality of Life and Well-Being Research* (p. 5986-5987). Springer, Cham. https://doi.org/10.1007/978-3-031-17299-1_2503

Smits, Y. (2023). A discursive analytical approach to understanding target audiences: How NATO can improve its actor-centric analysis. *The Hague Centre for Strategic Studies.*

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes, 62*(2), 159-174.

Soesanto, S. (2024). *The Ukrainian Way of Digital Warfighting: Volunteers, Applications, and Intelligence Sharing Platforms.* Center for Security Studies (CSS), ETH Zürich.

Soll, J. B., Milkman, K. L., & Payne, J. W. (2015). A user's guide to debiasing. *Wiley Blackwell Handbook of Judgment and Decision Making*, 924-951.

Sostaric, M. (2019). The American Wartime Propaganda During World War II. *Australasian Journal of American Studies, 38*(1), 17-44.

Soygeniş, İ. H. (2014). NATO adaptation and terrorism, Al-Qaïda security implication to the Balkan region. *4th International Conference on European Studies*. dspace.epoka.edu.al/handle/1/990

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review, 34*(3), 150-160. https://doi.org/10.1177/0266382117722446

Stewart III, C. (2022). Trust in elections. *Daedalus, 151*(4), 234-253.

Stoianov, N. (2021). Chapitre 12 – Conclusion générale et perspectives – La guerre cognitique et ses implications pour le panel IST de la STO. Dans B. Claverie, B. Prébot et F. Du Cluzel (dir.), *La Guerre Cognitique* (p. 12.1). CSO

Strzelecki, A., & Rizun, M. (2022). Consumers' Change in Trust and Security after a Personal Data Breach in Online Shopping. *Sustainability, 14*(10), 10. https://doi.org/10.3390/su14105866

Su, C., Zhou, H., Gong, L., Teng, B., Geng, F., & Hu, Y. (2021). Viewing personalized video clips recommended by TikTok activates default mode network and ventral tegmental area. *NeuroImage, 237*, 118136. https://doi.org/10.1016/j.neuroimage.2021.118136

Sun Tzu. (V^{e} siècle). *L'art de la guerre*.

Sunstein, C. (1999). The Law of Group Polarization. *Law & Economics Working Papers.* chicagounbound.uchicago.edu/law_and_economics/542

Surowiecki, J. (2004). *The Wisdom of crowds*. Doubleday; Anchor.

Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. *Cognition, 133*(3), 572-585. https://doi.org/10.1016/j.cognition.2014.08.006

Takagi, K. (2022). *The Future of China's Cognitive Warfare: Lessons from the War in Ukraine*. War on the Rocks. Consulté le 16 décembre 2022, à l'adresse https://warontherocks.com/2022/07/the-future-of-chinas-cognitive-warfare-lessons-from-the-war-in-ukraine/

Takeuchi, H., Taki, Y., Asano, K., Asano, M., Sassa, Y., Yokota, S., Kotozaki, Y., Nouchi, R., & Kawashima, R. (2018). Impact of frequency of internet use on development of brain structures and verbal intelligence: Longitudinal analyses. *Human Brain Mapping, 39*, 4471-4479. https://doi.org/10.1002/hbm.24286

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors We Think With: The Role of Metaphor in Reasoning. *PLOS ONE, 6*(2), e16782. https://doi.org/10.1371/journal.pone.0016782

Thompson, J. D., & Weldon, J. (2022). Audiences and Target Audiences. Dans *Content Production for Digital Media: An Introduction* (p. 11-20). Springer. https://doi.org/10.1007/978-981-16-9686-2_2

Trabelsi, B. (2023). *La guerre cognitive : Note de recherche*. Ministère des Armées. www.terre.defense.gouv.fr/sites/default/files/ccf/20231016_NP_CDEC-PEP-OC_Note-de-recherche-La-guerre-cognitive.pdf

Trilateral Research. (2025, janvier 24). Using Responsible AI to combat misinformation. trilateralresearch.com/responsible-ai/using-responsible-ai-to-combat-misinformation

Tsikhelashvili, N. (2022, mai 5). Influence of Cognitive Warfare on National Will to Fight. *The Defence Horizon Journal*. tdhj.org/blog/post/cognitive-warfare-national-will-fight/

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science, 211*(4481), 453-458. https://doi.org/10.1126/science.7455683

Umar, R., Sunardi, & Fitriana, Y. B. (2017). Taxonomy of Decision Support System Based on Software and Calculation Method. *International Journal of Innovative Science and Research Technology, 2*(9), 206-211.

Untersinger, M., Reynaud, F., & Leloup, D. (2024, décembre 18). Des milliers d'influenceurs, dont des Français, approchés par des personnes proches du Kremlin pour diffuser de la propagande prorusse. *Le Monde.* www.lemonde.fr/pixels/article/2024/12/18/guerre-en-ukraine-des-milliers-d-influenceurs-dont-des-francais-approches-pour-diffuser-de-la-propagande-prorusse_6455494_4408996.html

Vaillancourt, T. (2021). Mental Resilience and Coping With Stress: A Comprehensive, Multi-level Model of Cognitive Processing, Decision Making, and Behavior. *Frontiers in Behavioral Neuroscience, 15.*

Valette, M. (2024). Guerre cognitive, culture et récit national. *Ingénierie cognitique, 7*(1), 6-12.

VIGINUM. (2024a). *DISARM - Tactiques, techniques et procédures.* github.com/VIGINUM-FR/DISARM-FR/blob/main/DISARM_vf.pdf

VIGINUM. (2024b). *Portal Kombat : Un réseau structuré et coordonné de propagande prorusse*. www.sgdsn.gouv.fr/files/files/20240212_NP_SGDSN_VIGINUM_RAPPORT-RESEAU-PORTAL-KOMBAT_VF.pdf

Voelker, T. A., Wooten, K. C., & Mayfield, C. (2011). Towards a network perspective on change readiness. *Academy of Management Proceedings, 2011*(1), 1-6. https://doi.org/10.5465/ambpp.2011.65870187

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Social Science, 359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Vroom, V. H., & Yetton, P. (1973). Leadership and Decision-Making. *Administrative Science Quarterly, 18*(4). https://doi.org/10.2307/2392210

Vuving, A. L. (2009). *How soft power works.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1466220

Wakefield, J. (2022, mars 18). Deepfake presidents used in Russia-Ukraine war. *BBC.* www.bbc.com/news/technology-60780142

Walker, A. C., Turpin, M. H., Meyers, E. A., Stolz, J. A., Fugelsang, J. A., & Koehler, D. J. (2021). Controlling the narrative: Euphemistic language affects judgments of actions while avoiding perceptions of dishonesty. *Cognition, 211*, 104633. https://doi.org/10.1016/j.cognition.2021.104633

Walsh, V., Desmond, J. E., & Pascual-Leone, A. (2006). Manipulating brains. *Behavioural Neurology, 17*, 131-134.

Waltzmann, R. (2022). The role of today's VRE and considerations for Cognitive Warfare. *NATO's ACT.*

Ward, A. F., Duke, K., Gneezy, A., & Bos, M. W. (2017). Brain Drain: The Mere Presence of One's Own Smartphone Reduces Available Cognitive Capacity. *Journal of the Association for Consumer Research, 2*(2), 140-154. https://doi.org/10.1086/691462

Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology, 53*(2), 134 144. https://doi.org/10.1002/asi.10016

Weick, K. E., & Sutcliffe, K. M. (2015). *Managing the Unexpected: Sustained Performance in a Complex World.* John Wiley & Sons.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548-573.

Weldon, A. (2021, octobre 6). *Bytes not bombs: Student team works with NATO to define, track cognitive warfare attacks*. Johns Hopkins University - The Hub. https://hub.jhu.edu/2021/10/06/cognitive-warfare-attacks/

Wen, T., Chen, Y., Syed, T. A., & Ghataoura, D. (2025). Examining communication network behaviors, structure and dynamics in an organizational hierarchy: A social network analysis approach. *Information Processing & Management, 62*(1). https://doi.org/10.1016/j.ipm.2024.103927

Whiteaker, J., Sam Konen, C. (2021). Chapter 11 - Cognitive Warfare: Complexity and Simplicity. Dans B. Claverie, B. Prébot and F. du Cluzel (dir.), *Cognitive Warfare* (p. 11.1-11.4). CSO

Wilde, G. (2024, mars 11). *Why Cyber Attacks on Ukrainians Aren't Working the Way Russia Expected*. Carnegie Endowment for International Peace. carnegieendowment.org/emissary/2024/03/why-cyber-attacks-on-ukrainians-arent-working-the-way-russia-expected

Williams, B. (2024, août 1). Target Audience Research: How to, Techniques & Examples. *Insight7.* insight7.io/how-to-target-audience-research

Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., & Dobra, A. (2015). Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data. *PLOS One, 10*(7). https://doi.org/10.1371/journal.pone.0133630

Wineburg, S., & McGrew, S. (2019). Lateral Reading and the Nature of Expertise : Reading Less and Learning More When Evaluating Digital Information. *Teachers College Record, 121*(11), 1-40. https://doi.org/10.1177/016146811912101102

Wunder, M. (2021). Chapitre 7 – Les narrations submergent le monde. Dans B. Claverie, B. Prébot et F. Du Cluzel (dir.), *La Guerre Cognitique* (p. 7.1-7.4). CSO

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes, 83*(2), 260-281.

Yazgan, A. M. (2025). The Problem of the Century: Brain Rot. *OPUS Journal of Society Research, 22*(2), 2. https://doi.org/10.26466/opusjsr.1651477

Yaziji, M., & Doh, J. (2009). *NGOs and Corporations: Conflict and Collaboration*. Cambridge University Press.

Yeh, Y. Y. (2021). The Strategic Deployments of China's Cognitive Warfare Under Xi Jinping. *Taiwan Strategists, 2*, 1-18.

Yi, X., Ma, Y., Li, Y., Xu, H., & Ma, J. (2025). Artificial intelligence facilitates information fusion for perception in complex environments. *The Innovation, 6*(4). https://doi.org/10.1016/j.xinn.2025.100814

Yu, C., Cassar, I. R., Sambangi, J., & Grill, W. M. (2020). Frequency-Specific Optogenetic Deep Brain Stimulation of Subthalamic Nucleus Improves Parkinsonian Motor Behaviors. *Journal of Neuroscience, 40*(22), 4323-4334. https://doi.org/10.1523/JNEUROSCI.3071-19.2020

Zheng, R., Ospina-Forero, L., & Chen, Y. (2024). Implications of social network structures on socially influenced decision-making. *DECISION, 51*(1), 85-103. https://doi.org/10.1007/s40622-024-00380-5

Zioga, T., Ferentinos, A., Konsolaki, E., Nega, C., & Kourtesis, P. (2024). Video Game Skills across Diverse Genres and Cognitive Functioning in Early Adulthood: Verbal and Visuospatial Short-Term and Working Memory, Hand–Eye Coordination, and Empathy. *Behavioral Sciences, 14*(874). https://doi.org/10.3390/bs14100874

Zucchi, K. *Why Facebook is Banned in China & How to Access It*. Investopedia. Consulté le 22 décembre 2022, à l'adresse www.investopedia.com/articles/investing/042915/why-facebook-banned-china.asp

Zylberberg, J. (1975). Le Chili, l'Amérique latine et les États-Unis. *Études internationales, 6*(4), 555-562. https://doi.org/10.7202/700609ar

Annexes

Annexe 1 - Étude de cas : les chatbots et les fake news sur Twitter

Cette étude de cas a été menée en 2023 dans le cadre du Diplôme Universitaire « Big Data et Statistique pour l'Ingénieur » (BDSI) dispensé au sein l'École Nationale Supérieure de Cognitique. Elle est basée sur le jeu de données Twibot-20, créé par Feng et al. (2021), qui rassemble diverses données datant de 2020 sur les profils publics de 11826 utilisateurs de Twitter.

Le sujet de ce projet est la détection de bots sur le réseau social X (Twitter). En effet, il est bien connu que de nombreux faux utilisateurs sévissent sur ce réseau, créés par des robots et postant automatiquement du contenu. Les objectifs des personnes créant des bots peuvent être variés, mais ils sont souvent considérés comme mal intentionnés, étant donné que ces utilisateurs robots se font passer pour des humains. S'ils sont crédibles, ils peuvent ainsi chercher à désinformer ou influencer leurs lecteurs ou *« followers »* (pour influencer des élections, l'image publique d'une entreprise ou d'un projet, etc.) en les inondant de messages, vrais ou faux (fake news) allant dans un certain sens, noyant le contenu inverse auquel ils pourraient être exposés. Leur efficacité et leur crédibilité sont largement améliorées par des outils d'intelligence artificielle comme Chat-GPT, qui permet d'écrire en un temps record des textes construits sur un sujet donné, ou encore les générateurs automatiques d'images et plus particulièrement de visages, qui donnent une identité à ces bots.

Cette problématique s'inscrit dans celle, plus large, de la polarisation de la pensée des populations liée aux réseaux sociaux. En effet, les bots, plutôt que de présenter un contenu varié à l'utilisateur, tendent à l'exposer à un contenu qui le conforte dans ses idées, sans le contrarier, afin d'augmenter le temps d'utilisation du réseau social par l'utilisateur (Kubin & von Sikorski, 2021).

L'objectif de ce projet est donc de chercher à comprendre les différences entre utilisateurs bots et humains dans ce jeu de données, et de mettre en place une méthode d'apprentissage supervisé pour chercher à déterminer si un utilisateur est un bot ou un humain.

1.1. Influence et manipulation sur les réseaux sociaux

1.1.1. Mécanismes cognitifs et émotionnels exploités pour optimiser la viralité

Les réseaux sociaux exploitent largement des ressorts émotionnels pour maximiser la viralité. Sur Twitter, la surprise et le dégoût sont des émotions identifiées comme particulièrement efficaces pour provoquer des re-tweets viraux. Ainsi, Cocron et Aronhime (2022) proposent de prioriser la vérification sur les contenus véhiculant ces émotions, tout en soulignant que ralentir la diffusion des fake news pourrait suffire à limiter considérablement leur impact.

Les biais cognitifs jouent également un rôle fondamental dans la propagation virale de l'information. Les manipulateurs les exploitent sciemment ; mais ces biais, tels que le conformisme, sont aussi naturellement activés par les dynamiques sociales en ligne (Devillers, 2021). Devillers donne l'exemple du jeu du dictateur : il s'agit d'une expérimentation menée auprès d'enfants à qui on propose de

donner un certain nombre de billes à l'autre, le nudge consistant à leur faire croire que les autres enfants en donnent plus pour activer un biais de conformisme. Cela illustre comment des biais peuvent être activés et utilisés par des bots pour orienter les comportements sur les réseaux sociaux.

1.1.2. Influence par les États

Les réseaux sociaux ne sont pas de simples espaces d'expression : ils sont devenus des outils de guerre cognitive, exploités à la fois par des États étrangers et par les institutions nationales. Selon Cook (2023), le Pentagone aurait bénéficié d'un passe-droit de la part de Twitter pour créer des comptes fictifs en arabe. Ces comptes avaient pour but de mener des opérations d'influence psychologique, relayant des messages pro-guerre au Yémen, anti-Iran ou *« affirmant que les frappes de drones américaines ne frappaient que des terroristes »*. Cook mentionne également des suppressions de comptes, *« non pas parce que ce qui a été dit était de la désinformation vérifiable, mais parce que les tweets franchissaient des lignes rouges politiques »*.

1.1.3. Exploitation des données personnelles

Les plateformes comme Facebook permettent de prédire, avec un certain degré de fiabilité, des informations sensibles sur leurs utilisateurs : orientation sexuelle, appartenance ethnique, convictions religieuses ou politiques, traits de personnalité, intelligence, addictions, etc. (Cocron & Aronhime, 2022). Ces données ouvrent la voie à une manipulation individualisée, rendant les campagnes d'influence encore plus ciblées et efficaces.

Par ailleurs, selon Devillers (2021), jusqu'à 15 % des comptes actifs sur Twitter sont des bots. Ces agents automatisés influencent les discussions, administrent des groupes ou gonflent artificiellement le nombre d'abonnés, facilitant la diffusion de certaines idées et biaisant la perception de leur popularité.

1.2. Problématique de la détection de bots sur les réseaux sociaux

La détection des bots sur les réseaux sociaux est donc une question primordiale pour la préservation de l'authenticité des informations diffusées sur les réseaux sociaux, et la protection des utilisateurs humains.

Ainsi, de nombreuses caractéristiques des bots sur Twitter ou autres ont été relevées par des chercheurs, afin d'aider les utilisateurs à les reconnaître (Feng et al., 2021, Feng et al., 2023) :

- de nombreux posts étalés sur de longues périodes de temps (les robots ne dorment pas) ;
- nom de compte souvent aléatoire, qui ne signifie rien ;
- nombre de langues dans lesquelles le compte publie (par exemple, si la personne parle couramment 10 langues) ;
- un manque de pertinences et d'originalité des posts, voire une répétition de posts avec le même contenu ;
- des posts contenant des liens vers des sites de phishing (arnaque en ligne) ou des publicités, voire des liens non valides ;
- les bots ont tendance à se suivre les uns les autres sur Twitter pour augmenter leur réseau et leur crédibilité, ce qui fait qu'ils suivent des utilisateurs généralement moins connus que les utilisateurs humains.

Une combinaison de ces différentes caractéristiques peut aider à identifier manuellement et individuellement un utilisateur robot, mais il est important d'automatiser ces méthodes.

Malheureusement, les algorithmes de détection de bots ont leurs limites, étant donné que les bots évoluent rapidement pour leur échapper. Par exemple, pour passer sous le radar des algorithmes comptant le nombre de posts par 24h (un nombre anormalement élevé trahissant souvent un bot), certains ont évolués pour devenir très actifs pendant une certaine période avant d'hiberner pendant longtemps, ramenant ainsi leur ratio de tweets par jour à un niveau normal. Un autre exemple est que certains bots ont évolué pour mélanger des contenus à visée d'influence parmi des tweets volés à d'autres utilisateurs, afin d'être plus difficiles à repérer grâce à des outils d'analyse de texte.

Il s'agit donc d'une lutte sans fin entre les développeurs de bots aidés par les progrès de l'intelligence artificielle, et les développeurs de solutions de détection. Ces derniers utilisent principalement 3 types de méthodes de détection : l'analyse des données utilisateur (c'est le cas sur lequel nous allons nous pencher), l'analyse du texte (contenu des tweets et des descriptions des utilisateurs), et l'analyse de la forme des graphes des relations entre utilisateurs (followers, amis…).

1.3. Description du jeu de données

1.3.1. Collection du jeu de données

Un des objectifs principaux de la création de ce jeu de données était de pallier l'absence de jeu de données contenant une certaine diversité de contextes : ainsi, afin de créer un jeu de données avec des utilisateurs variés, Feng et al. (2021) ont sélectionné des *« seed users »* (utilisateurs de base), soit des personnes disposant d'une certaine notoriété dans 4 domaines différents (politique, business, sport, loisirs), puis ont sélectionné tous les utilisateurs ayant une relation avec ceux-ci (following, followers, amis…), remontant ainsi de plusieurs niveaux.

1.3.2. Labellisation du jeu de données

Le jeu de données a été labellisé manuellement par crowdsourcing (les chercheurs ont fait appel au public afin d'avoir de nombreuses personnes labellisant ce jeu de données conséquent).

Chaque profil du jeu de données a été étudié par 5 personnes différentes non expertes, qui l'ont alors labellisé comme étant humain ou bot. En cas de désaccord entre ces personnes, les chercheurs responsables du projet vérifiaient l'utilisateur et lui affectaient un label.

Une partie des profils ainsi labellisés a été contrôlée par des spécialistes, ce qui a permis d'évaluer que cette labellisation présentait un taux d'erreur de 20%. Ce taux est élevé et il s'agit d'un défaut important du jeu de données. Il montre que la labellisation est une problématique importante pour la détection.

1.3.3. Type de données prises en compte

Le jeu de données contient des informations quantitatives et qualitatives sur les profils des utilisateurs considérés :

- l'identifiant unique de chaque utilisateur ;
- des données sur le profil utilisateur :

    - données qualitatives : son identifiant, son nom, sa localisation, sa description, son url, sa date de création, son fuseau horaire, sa langue, les différentes couleurs utilisées sur son profil, l'URL de son image de profil…
    - booléens : s'il est « protected », si la géolocalisation est activée, s'il est vérifié, si la traduction est activée, s'il a une image de fond, s'il a un profil par défaut, s'il a une image de profil par défaut…
    - données quantitatives : son nombre de followers, son nombre d'amis, le nombre de listes dans lesquelles il a été intégré (listes diffusant les tweets des utilisateurs sélectionnés), son nombre de tweets enregistrés comme favoris, le nombre de statuts qu'il a postés ;
- les 200 derniers tweets postés par l'utilisateur ;
- la liste des following et followers ;
- une liste avec au moins 1 domaine d'intérêt de l'utilisateur parmi « politics », « entertainment », « sports » et « business » ;
- « label » vaut 0 si l'utilisateur a été labellisé comme humain ou 1 s'il a été labellisé comme un bot.

Ce jeu de données permet donc de faire de la détection en utilisant les données du profil, en utilisant une analyse du langage dans les tweets postés par l'utilisateur, ou une analyse des graphes de relations entre les utilisateurs. Nous nous concentrons ici sur l'analyse des données du profil utilisateur uniquement.

1.4. Statistiques descriptives

1.4.1. Répartition des utilisateurs dans le jeu de données

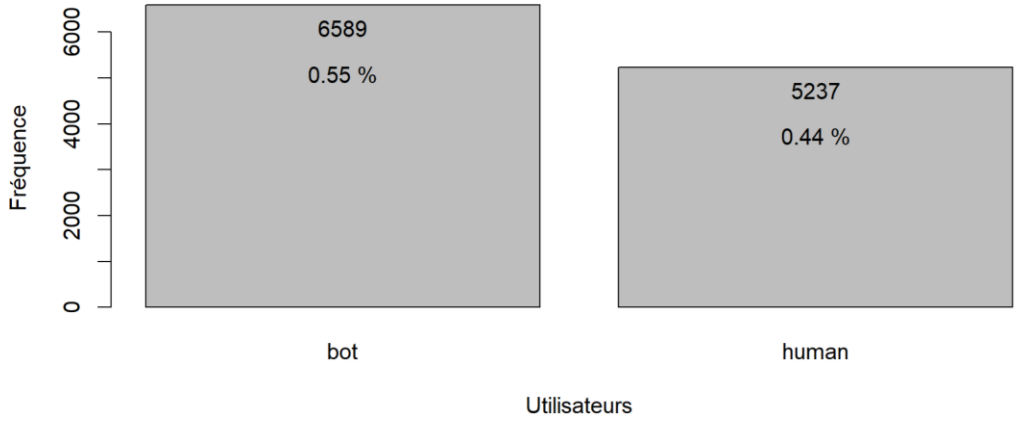

Figure 1 : Proportions d'utilisateurs bots et humains dans le jeu de données

La Figure 1 nous permet de voir que les utilisateurs bots sont plus nombreux dans le jeu de données (55%) que les utilisateurs humains. Or, Devillers (2021) mentionne qu'environ 15% des comptes actifs sur Twitter sont de faux comptes pilotés par des bots. Connaissant la manière dont les utilisateurs composant le jeu de données ont été récoltés en se basant sur les relations entre utilisateurs, nous pourrions imaginer que cette différence est due à la méthode de sélection (les graphes de relations des bots n'étant pas les mêmes que ceux des humains) ou à des erreurs de labellisation du jeu de données (20% d'erreur selon Feng et al.).

1.4.2. Répartition des données par domaine

Un des éléments qui nous intéressaient particulièrement était la répartition des utilisateurs par domaine (business, politique, loisirs, sports).

Les analyses suivantes sont faites sur une sous-partie du jeu de données contenant uniquement les 10175 utilisateurs n'ayant qu'un seul domaine d'activité (certains étant présents dans plusieurs domaines à la fois).

La Figure 2 montre qu'il y a plus d'utilisateurs bots qu'humains dans le domaine du sport, tandis que la répartition semble à peu près égale dans le domaine de la politique.

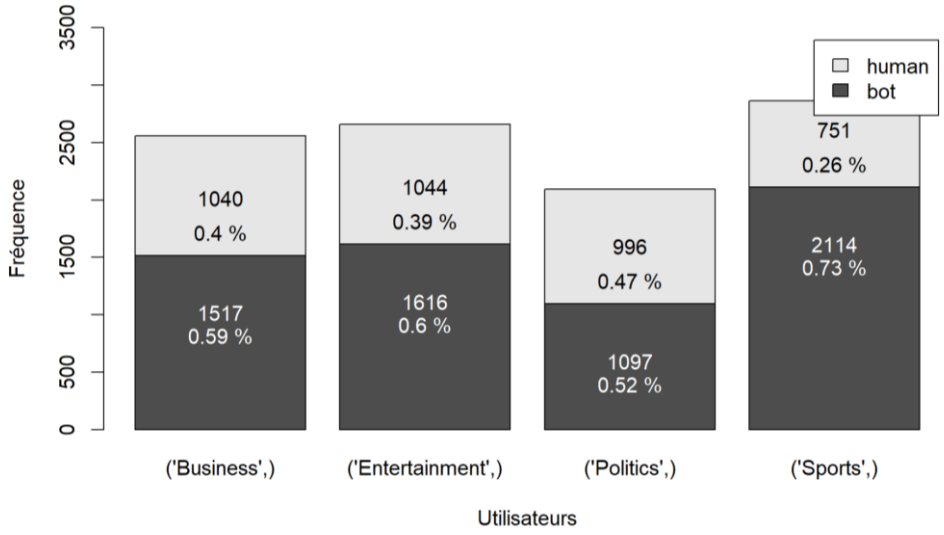

Figure 2 : Proportion d'utilisateurs bots et humains en fonction du domaine

L'AFC présentée en Figure 3 nous conforte dans l'idée qu'il y a plus d'utilisateurs bots dans le domaine du sport et plus d'humains dans le domaine politique, alors qu'ils sont à peu près répartis de manière équivalente dans les loisirs et le business.

Nous réalisons un test du chi-deux d'indépendance afin d'évaluer le lien entre les deux variables qualitatives « domaine » et « label ». La p-value obtenue est largement inférieure à 5%, il y a donc un fort rejet de l'hypothèse H0 qui postule qu'il y a une indépendance entre les variables domaine et label. Il est donc possible d'affirmer qu'il y a une dépendance importante entre le domaine d'activité et le label bot ou humain. Cependant, connaître le domaine d'activité d'un utilisateur n'est évidemment pas suffisant pour déterminer s'il s'agit d'un bot ou d'un humain.

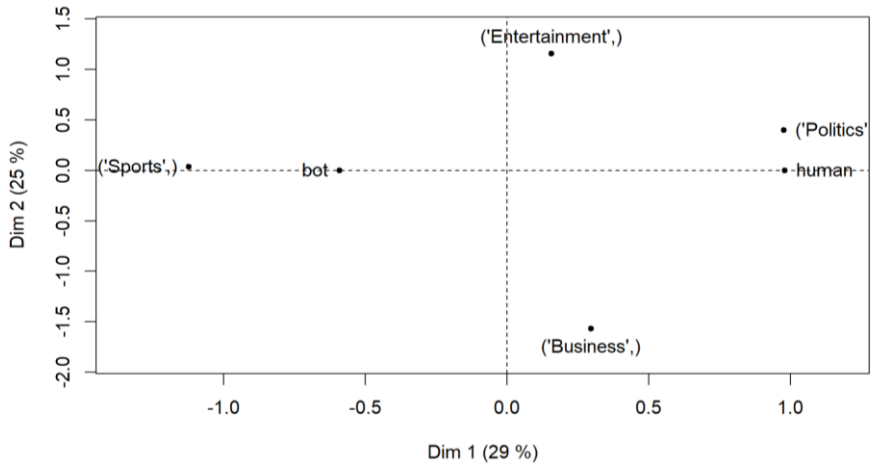

Figure 3 : AFC de la répartition des humains et bots suivant le domaine d'activité

1.4.3. Corrélations entre les variables

Nous étudions à présent les corrélations entre les variables quantitatives présentes : nombre de followers, nombre d'amis, nombre de listes dans lesquelles l'utilisateur a été intégré, nombre de tweets enregistrés comme favoris, nombre de statuts postés par l'utilisateur.

La Figure 4 et la Figure 5 montrent qu'il y a une corrélation entre le nombre de followers et le nombre de listes dans lesquelles l'utilisateur a été intégré, ce qui paraît logique : un utilisateur qui est suivi par de nombreux autres a une visibilité plus importante et est donc intégré à plus de listes, qui diffusent des tweets d'utilisateurs sélectionnés. Cette corrélation est de 80,4%. La coloration en fonction du label (en noir les utilisateurs humains, en rouge les bots) nous permet cependant de remarquer que cette corrélation concerne principalement les utilisateurs humains : les bots semblent avoir généralement peu de followers et être cités dans peu de listes.

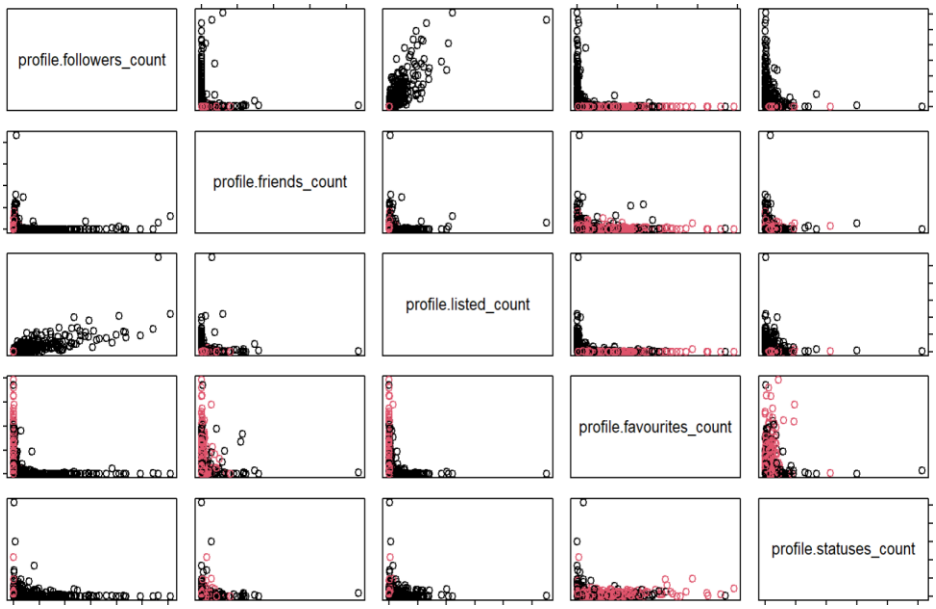

Figure 4 : Table des corrélations entre les variables quantitatives du jeu de données

Figure 5 : Matrice de corrélation entre les variables quantitatives du jeu de données

1.4.4. Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) permet de décrire, explorer et visualiser des données mixtes à la fois quantitatives et qualitatives (Chavent et al., 2014).

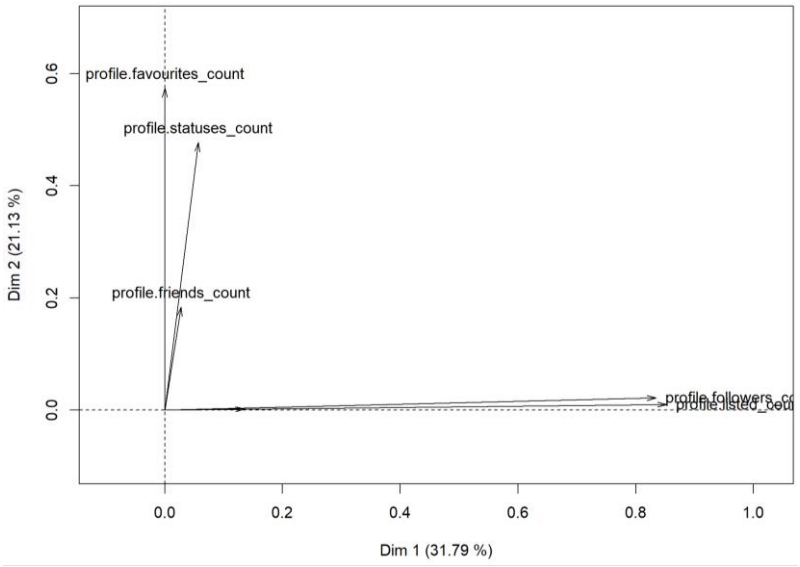

Figure 6 : Représentation ACP des données quantitatives

La représentation des données quantitatives avec les deux premières dimensions de l'ACP (Figure 6) contient 52,92% de l'information de ces variables. Nous retrouvons la corrélation entre le nombre de followers et la présence dans des listes, ainsi qu'une deuxième corrélation (moins significative) entre le nombre de favoris et le nombre de statuts. Le nombre d'amis semblerait corrélé également, mais il n'est pas représenté très significativement donc il est difficile de conclure à son sujet.

La représentation des données qualitatives avec les deux premières dimensions de l'ACP (Figure 7) ne contient que 19,13% de l'information de ces variables. Cela s'explique par le fait que ces variables sont moins corrélées, mais aussi plus nombreuses. Nous retrouvons toutefois le rapprochement entre la labellisation, la qualité vérifiée ou non du profil et son domaine d'activités.

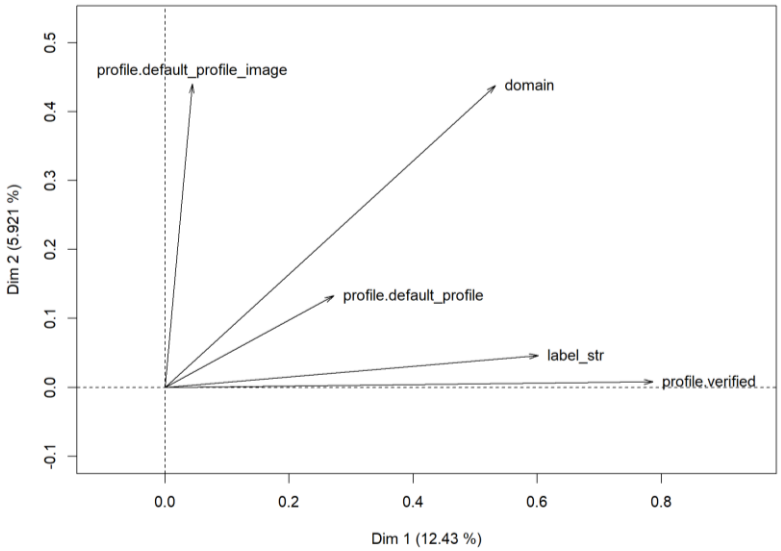

Figure 7 : Représentation ACP des données qualitatives

Sur la Figure 8, nous pouvons observer que la qualité de bot ou humain est mieux séparée sur les deux dimensions distinguées par l'ACP des variables quantitatives (à gauche), mais nous distingons tout de même deux zones dans la représentation suivant les deux dimensions distinguées par l'ACP des variables qualitatives (à droite). Cependant, ces zones se recouvrent dans les deux cas et ces dimensions ne suffisent donc pas à déterminer si un utilisateur est un bot ou un humain.

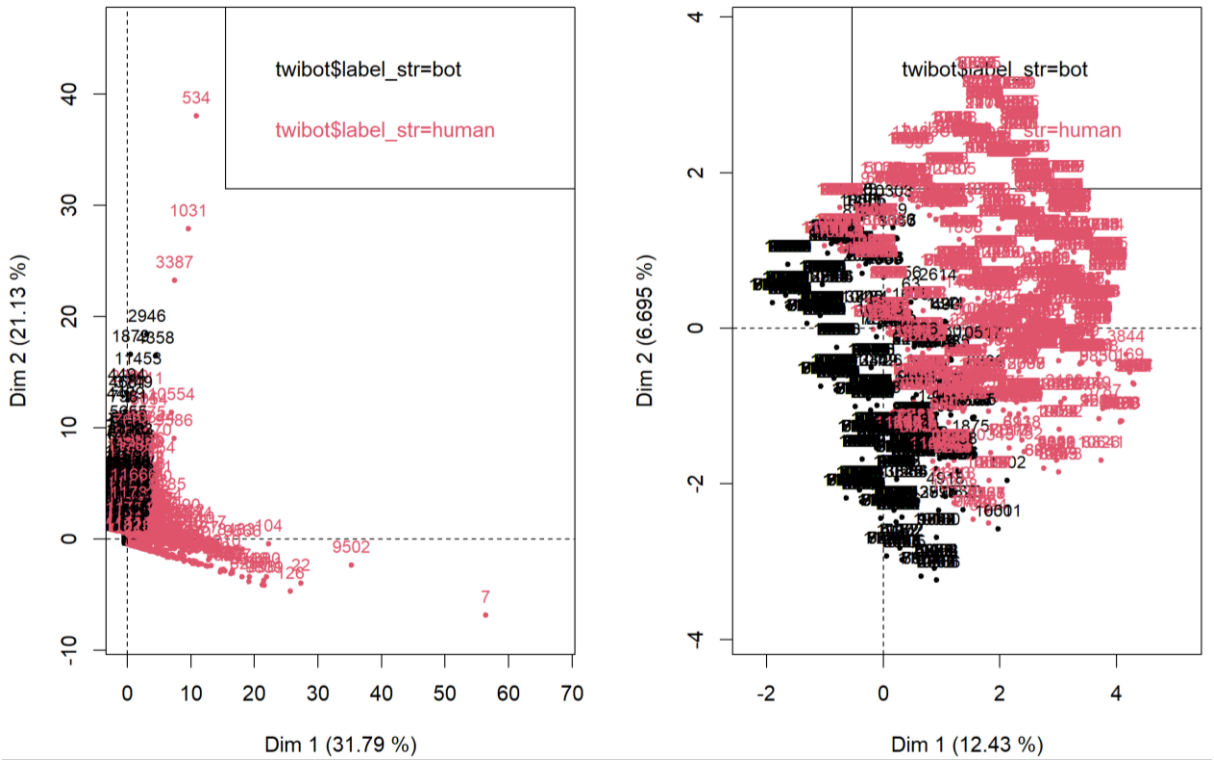

Figure 8 : Répartition des individus suivant les 2 premières dimensions des ACP quantitative et qualitative

1.5. Apprentissage supervisé

1.5.1. Règle de Bayes avec coûts

Un premier apprentissage supervisé avec l'algorithme LDA, utilisant la méthode des coûts de Bayes, est effectué uniquement avec les variables quantitatives du jeu de données. Le jeu de données est donc découpé en un échantillon de 80% des observations constituant le jeu de données d'apprentissage, et les 20% restants sont utilisés pour le test. Les proportions d'utilisateurs bots et humains dans ces deux échantillons sont équivalentes.

	pred_test	
Ytest	bot	humain
bot	1330	13
human	896	127

Figure 9 : Matrice de confusion des prédictions de l'algorithme LDA sans ajustement des coûts bayésiens

Un premier entraînement et prédictions effectuées sans ajuster les coûts bayésiens produit un taux d'erreur de 43,79%. Ce taux d'erreur est élevé, mais en étudiant la matrice de confusion (Figure 9), nous observons que 1330 bots sont correctement prédits (seuls 13 ont été prédits comme étant des humains) ; cependant, 896 humains ont été classés comme étant des bots (donc plus de 85% des humains sont prédits comme étant des bots). Nous pouvons supposer que cette prédiction s'est beaucoup appuyée sur la corrélation entre la labellisation et le nombre de followers ou d'intégrations à des listes : nous aurions alors détecté tous les utilisateurs ayant un faible nombre de followers et d'intégration à des listes (soit tous les utilisateurs ayant peu de notoriété) comme étant des bots.

Nous augmentons le coût de mauvaise classification d'un humain pour chercher à avoir une meilleure répartition des prédictions. Avec un coût de 4 pour la mauvaise classification d'un bot et de 5 pour la mauvaise classification d'un humain, nous obtenons un taux d'erreur de prédiction de 46,20%, ce qui est légèrement plus élevé qu'avec la fonction LDA ; c'est normal, puisque nous cherchons à minimiser le risque de mauvaise attribution, plutôt que le risque d'erreur.

	pred_test	
Ytest	bot	humain
bot	1273	70
human	693	330

Figure 10 : Matrice de confusion des prédictions de l'algorithme LDA avec ajustement des coûts bayésiens

La matrice de confusion alors obtenue (Figure 10) est légèrement mieux répartie, mais nous observons toujours que la majorité des humains sont détectés comme étant des bots ; si nous augmentons encore le risque, alors la majorité des humains sont détectés comme étant des humains, mais la majorité des bots sont également détectés comme étant des humains. Différents essais avec des coûts variés n'ont pas donné de résultats plus probants.

Nous pouvons donc estimer qu'utiliser uniquement les données numériques présentes dans le jeu de données ne permet pas de discriminer efficacement les humains et les bots. Nous tentons d'effectuer une prédiction via une régression logistique, afin de prendre en compte à la fois les variables numériques et les variables qualitatives.

1.5.2. Régression logistique

Un modèle de régression logistique est donc mis en place en utilisant la bibliothèque scikit-learn. Les variables utilisées pour cette régression linéaire sont toutes les variables quantitatives et booléennes (en omettant donc les variables de couleurs et les domaines d'intérêt). La variable « domaine » est également exclue pour simplifier le traitement et l'analyse ; elle aurait pu être intégrée grâce à un encodeur One Hot, car il s'agit d'une variable intéressante qui, dans les analyses précédentes, semblait liée à la labellisation.

Les données sont séparées en jeu de données et jeu de test, en isolant la variable « target » (le label).

Le modèle de régression logistique, testé sur le jeu de données de test, donne une précision de 0,67 dans ses prédictions de label (humain ou bot), soit un taux d'erreur de 33%. Il s'agit donc d'un progrès significatif comparé aux 44% d'erreur de l'algorithme LDA (paragraphe *1.5.1*). Étant donné que le jeu de données comporte 20% d'erreur de validation, ce résultat semble cohérent et acceptable, bien qu'il ne permette pas une détection particulièrement sûre.

1.6. Discussion

Cette analyse de données a été ciblée sur les données du profil des utilisateurs de Twitter. Nous avons vu que certaines variables étaient corrélées avec le label bot ou humain, mais avec un biais important : par exemple, la variable du nombre de followers, qui permet d'identifier facilement les utilisateurs humains disposant d'une forte notoriété (célébrités), mais ne permet pas de différencier les utilisateurs peu actifs des robots.

De même, la mise en place d'un apprentissage supervisé avec des coûts bayésiens n'a pas donné des résultats de prédiction probants. Seule la régression logistique nous a permis de discriminer de manière satisfaisante les utilisateurs bots et les utilisateurs humains. Son efficacité de 67% est relative, mais doit être remise en perspective en rappelant les 20% d'erreur dans la labellisation du jeu de données, qui produisent un certain bruit dans les données. Ce projet a donc permis de démontrer la possibilité de détecter, dans une certaine mesure, les bots sur Twitter en utilisant les données du profil des utilisateurs.

Il faut également rappeler que les bots évoluent, et comme le démontrent Feng et al. (2021), chaque algorithme de détection de bots entraîné sur un certain jeu de données voit ses performances baisser sur des jeux de données plus récents. Pour effectuer une détection automatique de bots sur Twitter, il faudrait donc se baser sur un jeu de données très récent, labellisé de manière plus efficace, et éventuellement utiliser d'autres méthodes (analyse du texte des tweets et de la structure des relations entre utilisateurs) pour s'assurer d'une détection plus précise et ainsi éviter d'exclure des utilisateurs humains. Il s'agit d'un problème ouvert, en constante évolution aussi bien au niveau de sa complexité que des solutions proposées.

Il est important de rappeler que les bots ne sont pas nécessaires pour rendre un message viral : les réactions de surprise et de dégoût suffisent pour que les vrais utilisateurs les repartagent en masse (Cao et al., 2021).

Il existe d'autres méthodes de détection de bots mais aussi de désinformation, efficaces et systématiques : par exemple, la détection automatique de textes contradictoires, qui *« permet d'identifier à partir de textes de référence – la "vérité" – tous les messages sur les réseaux sociaux qui cherchent à contredire des faits »*, ou encore le topic modeling qui *« permet de regrouper*

automatiquement… les documents traitant d'un même sujet… » et ainsi *« d'identifier de nouveaux sujets le plus tôt possible, en particulier afin d'identifier des anomalies »* (Auguste & Bresson, 2024).

1.7. Bibliographie (annexe 1)

Auguste, J., & Bresson, E. (2024). Neutraliser la désinformation en temps réel avec l'IA. Dans *Lutte contre les manipulations de l'information—Regards croisés de spécialistes et d'acteurs du domaine* (Pôle d'Excellence Cyber, Vol. 2). https://www.pole-excellence-cyber.org/evenements/lutte-contre-les-manipulations-de-linformation-decouvrez-le-tome-2/

Cao, K., Glaister, S., Pena, A., Rhee, D., Rong, W., Rovalino, A., Bishop, S., Khanna, R., & Singh Saini, J. (2021, mai 20). Countering cognitive warfare: Awareness and resilience. *NATO Innovation Hub.* https://www.nato.int/docu/review/articles/2021/05/20/countering-cognitive-warfare-awareness-and-resilience/index.html

Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2014). *Multivariate Analysis of Mixed Data: The R Package PCAmixdata* (arXiv:1411.4911). arXiv. http://arxiv.org/abs/1411.4911

Cocron, A., & Aronhime, L. (2022). *Cognitive Warfare: What's Next?* NATO Cognitive Warfare Workshop - USMA West Point.

Cook, J. (2023). *Comment les réseaux sociaux sont devenus une « filiale » du FBI et de la CIA.* Middle East Eye édition française. http://www.middleeasteye.net/fr/opinionfr/etats-unis-twitter-files-manipulation-reseaux-sociaux-fbi-cia-pentagone-influence-politique

Devillers, L. (2021). Désinformation : Les armes de l'intelligence artificielle. *Pour la science, 523*. https://www.pourlascience.fr/sd/informatique/desinformation-les-armes-de-l-intelligence-artificielle-21678.php

Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., Zheng, Q., Zhang, W., Lei, Z., Yang, S., Feng, X., Zhang, Q., Wang, H., Liu, Y., Bai, Y., Wang, H., Cai, Z., Wang, Y., Zheng, L., … Luo, M. (2023). *TwiBot-22: Towards Graph-Based Twitter Bot Detection* (arXiv:2206.04564). arXiv. http://arxiv.org/abs/2206.04564

Feng, S., Wan, H., Wang, N., Li, J., & Luo, M. (2021). TwiBot-20: A Comprehensive Twitter Bot Detection Benchmark. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4485-4494. https://doi.org/10.1145/3459637.3482019

Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association, 45*(3), 188 206. https://doi.org/10.1080/23808985.2021.1976070

Annexe 2 - Grilles de jeu des parties A, B, C de la phase C de l'expérimentation 2

Lors des 3 parties jouées à la phase C, *Léa* et *Thomas* sont simulés par un expérimentateur *agent*. La première colonne contient la météo de chaque tour, qui détermine combien d'eau les joueurs peuvent récolter. Dans la colonne « Participant » apparaissent les cartes piochées par le joueur à chaque tour, avec en gras les choix les plus significatifs concernant sa stratégie. Dans les colonnes « Léa » et « Thomas », les actions à faire par l'expérimentateur apparaissent en rouge et les cartes à conserver apparaissent en vert (il devra jeter les autres). Les parties sont jouées dans un ordre différent pour chaque participant afin de contrebalancer les effets d'apprentissage, de fatigue ou de l'influence tentée au cours de l'expérimentation (2e et 3e partie).

PARTIE A	**Participant**	**Léa**	**Thomas**
Main initiale	Poisson pourri, somnifères, taser, sandwich	Carte inutile, gourde, sandwich, bouteille d'eau	Carte inutile, bouteille d'eau, revolver, panier garni
Tour 1 - météo 1	Moulin à légumes	Carte inutile	**Noix de coco**
Tour 2 - météo 2	Carte inutile	Carte inutile / **pose gourde**	Carte inutile
Tour 3 - météo 0	**Poupée vaudou**	Carte inutile	Carte inutile / **pose le revolver + info "si on a vraiment besoin d'éliminer quelqu'un pour survivre"**
Tour 4 - météo 3	Carte inutile	Bouteille d'eau	Carte inutile
Tour 5 - météo 0	Sandwich	Carte inutile	**Planche (poser)**
Tour 6 - météo 1	Nouilles chinoises	**Sardines**	Carte inutile
Tour 7 - météo 2	**Cartouche**	Carte inutile	**Hache (poser)**
Tour 8 - météo 1	Bouteille d'eau	Carte inutile	Carte inutile
Tour 9 - tempête	Carte inutile	Carte inutile	**Kit BBQ Cannibale**

PARTIE B	Participant	Léa	Thomas
Main initiale	Cartouche, anti-venin, nouilles chinoises, bouteille d'eau	Plaque de tôle, noix de coco, bouteille d'eau, **conque** (utiliser si vote)	Plaque de tôle, moulin à légumes, sandwich, carte inutile
Tour 1 - météo 0	Sandwich	**Panier garni**	Carte inutile / **Thomas malade** (pioche de bois risque négatif)
Tour 2 - météo 2	Café	Carte inutile	**Sardines**
Tour 3 - météo 2	**Revolver**	Carte inutile	Carte inutile
Tour 4 - météo 0	Carte inutile	**Canne à pêche** (poser)	Carte inutile
Tour 5 - météo 1	Planche	Carte inutile	Carte inutile
Tour 6 - météo 3	Eau croupie	**Gourde** donner à Thomas	Carte inutile / **remercier gourde + chercher 6 eau**
Tour 7 - météo 1	**Somnifères**	Carte inutile	Carte inutile
Tour 8 - tempête	Carte inutile	Carte inutile / **poser conque si pas utilisée**	Carte inutile

PARTIE C	Participant	Léa	Thomas
Main initiale	Sandwich, kit BBQ Cannibale, plaque de tôle, bouteille d'eau	**Hache**, sandwich, sandwich, bouteille d'eau	Taser, moulin à légumes, longue-vue, bouteille d'eau
Tour 1 - météo 2	Carte inutile	Carte inutile / **poser hache**	Carte inutile
Tour 2 - météo 2	Poupée vaudou	Carte inutile	Carte inutile / **utiliser Longue-vue sur participant**
Tour 3 - météo 0	**Longue-vue**	**Sardines**	Carte inutile
Tour 4 - météo 1	Carte inutile	Carte inutile	Carte inutile
Tour 5 - météo 3	Gourde	Carte inutile	Carte inutile / **Taser : voler gourde sinon hache**
Tour 6 - météo 0	**Conque**	Carte inutile	Carte inutile
Tour 7 - météo 1	Anti-venin	**Noix de coco**	Carte inutile
Tour 8 - tempête	Sandwich	Carte inutile	Carte inutile

Annexe 3 - Liste des productions scientifiques associées à cette thèse

Communications lors de colloques avec actes :

Morelle, M., Marion, D., Cegarra, J., André, J-M. (2023). Le cognitive warfare : contexte, concept et enjeux*. 12ème Colloque de Psychologie Ergonomique - Épique 2023*, 434-438. https://arpege-recherche.org/user/pages/06.activites/03.colloques-epique/13.12e-colloque-epique/EPIQUE%202023_Paris-EVDG_Actes%20du%20Colloque.pdf

Morelle, M., Marion, D., Cegarra, J., & André, J.-M. (2023). Towards a Definition of Cognitive Warfare. *Actes de la conférence CAID 2023*, 37-40. https://hal.sorbonne-universite.fr/INUC/hal-04328461v1

Posters :

Cognitive Warfare : conception d'un système technologique permettant les accords de travail en équipes Humains / Machines pour la prise de décision en gestion de crise. Poster présenté lors de la *Journée de l'École Doctorale SPI*, (Bordeaux, France, 15 février 2024).

Morelle-Gerritsen, M., Marion, D., Cegarra, J., Unrein, H., Letouzé, T., & André, J.-M. (2025). Evaluating and Influencing Strategy in Real-time: Example of a Collaborative Strategy Game. Dans I. Wiafe, A. Babiker, J. Ham, K. Oyibo, & E. Vlahu-Gjorgievska (Éds.), *Persuasive Technology. PERSUASIVE 2025 Satellite Events.* Springer Nature Link. https://link.springer.com/book/9783031971761

Autres contributions :

Participation au Workshop « Guerre cognitive : points de vue et controverses » le 29/04/2024 à l'INALCO en compagnie de nombreux experts de la guerre cognitive dans le cadre du groupe CIVIL du projet GECKO (ENSC, INALCO, EGE).

Morelle, M. (2024). Que peut apporter la notion de Guerre Cognitive à la Lutte contre les Manipulations de l'Information ? Dans *Lutte contre les manipulations de l'information. Regards croisés de spécialistes et d'acteurs du domaine* (Vol. 2, p. 14-15). Pôle d'Excellence Cyber. https://www.pole-excellence-cyber.org/wp-content/uploads/2024/11/LMI_Tome2_PEC2024.pdf

Index nominum

AID : Agence de l'Innovation de Défense, organisme français de financement, de coordination et de valorisation de la recherche en innovation militaire. Organisme co-financeur de cette thèse.

ANTICIPE : Système d'aide à la décision développé par THALES, destiné à soutenir la prise de décision stratégique, notamment en détectant les informations critiques (CCIR – Commander's Critical Information Requirements) et en proposant des solutions adaptées au contexte détecté.

ByteDance : Entreprise technologique chinoise propriétaire des applications TikTok et DouYin, spécialisée dans les algorithmes de recommandation et la diffusion de contenus numériques.

Cambridge Analytica : Société britannique de conseil politique, connue pour avoir exploité les données d'utilisateurs de réseaux sociaux à des fins de ciblage électoral et de manipulation de l'opinion.

Delta : Plateforme numérique de commandement et de renseignement développée par l'armée ukrainienne, permettant la collecte, la visualisation et le partage en temps réel d'informations tactiques sur le champ de bataille.

DouYin : Version chinoise de TikTok, soumise à une régulation des contenus plus importante que ce dernier.

EGE : École de Guerre Économique, institution française d'enseignement supérieur spécialisée dans l'intelligence économique et la stratégie d'influence.

ENSC : École Nationale Supérieure de Cognitique, grande école d'ingénieurs française rattachée à Bordeaux INP, spécialisée en ingénierie cognitique, ergonomie, interaction homme-système et sciences cognitives appliquées.

Facebook : Réseau social en ligne fondé en 2004, permettant le partage de contenus, la création de communautés et la diffusion d'informations auprès d'un large public.

Galèrapagos : Jeu de société semi-collaboratif dans lequel les joueurs, naufragés sur une île déserte, doivent récolter et gérer des ressources afin de survivre et s'enfuir de l'île.

HFM ET-356 : Groupe de recherche Human Factors and Medicine Exploratory Team 356, notamment à l'origine du House Model sur la guerre cognitive.

Maven (Projet Maven) : Programme du Département de la Défense des États-Unis visant à intégrer l'intelligence artificielle dans l'analyse d'images et la reconnaissance automatique de cibles à des fins militaires.

NAFO : North Atlantic Fella Organization, communauté en ligne qui lutte contre la désinformation par l'humour.

Netflix : Plateforme de diffusion de films et séries en streaming.

Neuralink : Entreprise américaine fondée par Elon Musk, spécialisée dans le développement d'implants cérébraux et d'interfaces cerveau-ordinateur.

OTAN (NATO) : Organisation du Traité de l'Atlantique Nord, alliance politico-militaire fondée en 1949 pour assurer la défense collective des pays membres.

Palantir Technologies : Entreprise américaine spécialisée dans l'analyse de données massives (big data) et les solutions logicielles de renseignement pour les secteurs gouvernemental et industriel. Contributeur du projet Maven.

Panadarchipel : Application numérique conçue dans le cadre de cette thèse pour reproduire le jeu Galèrapagos dans un environnement expérimental, avec adaptation des règles au protocole d'étude (expérimentation 2).

Postdare : Outil logiciel expérimental d'aide à la décision inspiré du système ANTICIPE, développé dans le cadre de cette thèse pour la détection et l'analyse en temps réel de la situation et la proposition de contre-mesures.

Portal Kombat : Réseau de sites de désinformation pro-russe identifié par VIGINUM en 2023, diffusant des contenus manipulés en Europe et en Afrique.

Team Jorge : Officine israélienne spécialisée dans la désinformation sur les réseaux sociaux, accusée d'ingérence numérique et de manipulation de l'opinion publique à travers de fausses identités et des campagnes coordonnées sur les réseaux sociaux.

THALES : Groupe industriel français spécialisé dans la défense, la sécurité, l'aéronautique et les systèmes d'information critiques, financeur de cette thèse et concepteur et propriétaire du système ANTICIPE.

TikTok : Réseau social chinois de partage de vidéos courtes, propriété de ByteDance, largement utilisé pour le divertissement et la diffusion de contenus viraux.

VIGINUM : Service français de vigilance et de protection contre les ingérences numériques étrangères, placé sous l'autorité du Premier ministre, chargé de la détection et de la lutte contre la désinformation.

X (Twitter) : Réseau social fondé en 2006 (anciennement Twitter), basé sur la publication de messages courts et l'interaction en temps réel entre utilisateurs.