Mapping AI Issues in Media Through NLP Methods

Maxime Crépel[1], Salomé Do[1,2], Jean-Philippe Cointet[1], Dominique Cardon[1] and Yannis Bouachera[3]

[1]*médialab, Sciences Po Paris, 84 rue de Grenelle, 75007 Paris, France*
[2]*LATTICE, CNRS & École Normale Supérieure/PSL & Univ. Sorbonne nouvelle, France*
[3]*ENSAE Paris, 5 Avenue Le Chatelier, 91120 Palaiseau, France*

Abstract

Using a variety of NLP methods on a corpus of press articles, we show that there are two dominant regimes of criticism of artificial intelligence that coexist within the media sphere. Combining text classification algorithms to detect critical articles and a topological analysis of the terms extracted from the corpus, we reveal two semantic spaces, involving different technological and human entities, but also distinct temporality and issues. On the one hand, the algorithms that shape our daily computing environments are associated with a critical discourse on bias, discrimination, surveillance, censorship and amplification phenomena in the spread of inappropriate content. On the other hand, robots and AI, which refer to autonomous and embodied technical entities, are associated with a prophetic discourse alerting us to our ability to control these agents that simulate or exceed our physical and cognitive capacities and threaten our physical security or our economic model.

Keywords

algorithms, controversy, semantic network, ethics, AI, NLP, ML

1. Introduction

This article investigates how the press engages critically with Artificial Intelligence (AI). We leverage natural language processing algorithms to highlight the many ways that English speaking media frames certain AI applications as a public problem. We are especially interested in uncovering the various type of criticisms actors and institutions (e.g. scientific experts, civil society, companies, etc.) involved in the various applied domains where AI is a source of public concern [23, 17]. AI serves as an umbrella term that may refer to a more or less complex set of computer techniques ranging from simple automated algorithms to more complex deep learning machine learning systems. In order to give an operative and agnostic definition of AI, we will coin AI as a ”computing agent” considering all the techniques that, through automated computation, produce an output from any type of data. This broad definition allows us to understand how the actors themselves define these technologies and connect them to a plurality of socio-technical entities [41].

AI-powered technologies trigger debates in the public space every time a new application makes the headlines. These debates are most often organized around a clear-cut polarity

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

maxime.crepel@sciencespo.fr (M. Crépel); salome.do@sciencespo.fr (S. Do); jeanphilippe.cointet@sciencespo.fr (J. Cointet); dominique.cardon@sciencespo.fr (D. Cardon); yannis.bouachera@ensae.fr (Y. Bouachera)
0000-0001-9032-6538 (M. Crépel); 0000-0002-6095-6253 (S. Do); 0000-0002-3172-3960 (J. Cointet); 0000-0002-2384-5677 (D. Cardon)

opposing their risks and promises. Previous research [38, 6, 16] analyzing the media has shown an increasing visibility of the AI topic in media coverage in recent years, focusing mostly on technological advances in the field and on the potential developments of these technologies in business and industry. Ethical issues and social problems such as the risks of discrimination, bias or privacy are underrepresented in the media space. However, when the media sphere stresses the controversies on the negative effects of these technologies, it fosters a critical discourse towards AI, that contributes to building public opinion but also defines the forms of acceptability of these technologies in our societies. In this way, the media space seems to be the place where different types of critical discourses about AI are made visible.

First, the media reports the discussions and decisions that are made on ethical principles and modes of regulation of these technologies produced by experts and policy-makers. These ethical principles focus mainly on accountability, fairness and explainability but often suffer from a very broad level of generality. As a consequence, applying them to specific contexts can prove challenging [32, 18].

Second, the media also provides a coverage of the debates highlighted by academic research which focus on the difficulties of understanding how algorithmic systems work, and the control (or lack thereof) of data access by companies developing these technologies. This research points to the issues of transparency, loyalty and privacy [36], but it also denounces the capacities of those technologies to produce inequalities and discriminations [35, 39]).

Finally, the media gives visibility to cases that concern ordinary users. These users often denounce problematic situations they encounter in their daily use of these technologies. They then wonder how they are calculated and are also interested in the data sources that feed these calculation devices: [7].

Our goal is to capture those three types of reflexive and critical discourses around AI. To do so, we use Natural Language Processing algorithms on a specialized corpus of press articles to characterize the structure of the critical discourses about AI in the media sphere. More precisely, we aim at answering a series of interrelated questions that sheds light on this critical discourse. What are the main issues these technologies trigger? What types of actors and technical entities are involved in those controversies? Which disorders are these technologies being accused of?

2. Related Work

From a method perspective, this paper belongs to the larger field of computational sociology [14] with a special focus on text analysis. Texts are considered as data points which mining can help to illuminate social phenomena [20, 15]. Numerous types of computational methods related to text analysis have recently been developed or adapted to social science inquiries such as topic modeling [3], sentiment analysis [28] or word embedding [31]. Social sciences scholars have also been using such techniques to investigate the structure, framing and tone of press article databases [13].

Quantitative news related to content research is an historical topic in sociology and political science: it dates back to Weber who projected to using the content published by the press to monitor public opinion (see [27]) which was later operationalized by the Columbia school [26]. The availability of massive datasets of news articles online, combined with the development of new algorithms capable of "summarizing" and "mapping" the structure of those large text collections has contributed to a renewed interest in quantitative analyses in media studies.

Various questions can traditionally be addressed at a systemic level concerning the structure of the media ecosystem [2], the economy of the press outlets [29] or the way the media agenda is negotiated in-between actors [1]. Here we adopt a more "localized" approach, using NLP techniques to describe the variety of grievances addressed by the press to AI.

Other articles are using press articles to reveal the public perception of AI. They usually have a local focus (see [6] on the British press, [45] on the Dutch press, [12] focusing on the US coverage of AI or [34] focusing on two magazines). More importantly they are mainly concerned with describing the large trends accompanying the emerging technology. In this work we scrutinize in depth the nature and operators of the criticisms addressed to AI. In a previous research project [9] based on a qualitative approach of 50 mediated cases of algorithms-related problems, we analyzed the critical discourses in the media about those technologies. By analyzing the actantial system in these articles,[1] we have shown what kind of difficulties computing technologies produce on society, the multiplicity of the causes and the underlying principles of justice mentioned by the victims engaged in those issues. The objective of this research is to test whether such qualitative models can scale up and be applied to the processing of large corpora of articles. Our work lies at the intersection of traditional method approaches inspired from pragmatist theory and the sociology of critique [5] and recent modeling techniques originating from computer sciences and the larger field of AI. We also argue that our modeling strategy is also germane to attempts by other scholars in sociology of culture who are trying to identify (from within the text) the complex "role structure" played by certain actors and entities in press articles [33, 44].

3. Methodology

3.1. Dataset

Our corpus was extracted from a carefully curated lexical query[2] on AI and related techniques on a set of 47 generalist English-written press sources (27 sources from the US and 20 sources from the UK) available on the press platform Factiva.[3] It is composed of 29 342 press articles and spans over 5 years (from 2015 and onward).

The volume of articles extracted from our query increases over the 5-year time period, showing that the topic of algorithms and AI has gained momentum in the media in recent years. To control for the possibility that we are observing a general growth of the press article database, we compared the time evolution with collections observed from other queries ran over the same period and the same sources, for which a stable distribution is expected. The observation of the volume of articles retrieved from our query shows a significant increase (+163%) while articles dealing with the queries art (-9%), culture (+2%), economics (+0%) or even technology (+14%) remain relatively stable.

[1] Actantial system here refers to the idea coming from pragmatic sociology that both human and non-human entities of any scale can participate to the definition of the problem at stake [25].

[2] `"artificial intelligence" OR "AI" OR "algorithm*" OR "machine learning" OR "deep learning" OR "neural network*" NOT ("amnesty international" OR "weiwei" OR "air india") - We excluded a few words from the query because of the noise they produce on the final results. The first name of the famous artist "Weiwei" is ``Ai'' and the acronyms of Amnesty International and Air India are also misleading.`

[3] https://professional.dowjones.com/factiva/

3.2. Supervised classification of critical articles

Our press articles corpus is mostly composed of non-controversial content about AI. Newspapers may comment on the latest web innovation from web companies, discuss in a neutral tone the consequences of IoT for businesses, or comment on AI-powered predictions about climate disasters. Only sometimes, AI related computing agents are questioned and suspected explicitly. To map the critical discourses around AI, we then need to first identify those occurrences. We do not pretend our corpus to be composed of a comprehensive list of critical articles published by the press. The selection of sources is contingent to the data repository we use (Dow Jones Factiva) and our criterion to filter in articles is strict. We target articles which title explicitly contains elements of critique against AI. Article's title offers a good vantage point to judge the overall of the journalist toward AI. We think this is a reasonable choice as these articles are likely to be the ones containing the most articulated critical discourse. Additionally, the annotation and the training of the classifier is simpler and more efficient when considering shorter snippets of texts than the article lede or full text.

First, we start by manually annotating 6 257 article titles explicitly featuring a criticism toward any form of computing agent (algorithm, AI, robot, etc.).[4] Articles with neutral, positive or ambiguous statements or which do not directly refer to AI or computing agents are annotated as non-critical. For instance, "*AI to create more than 7M jobs*", "*In the 2020s, artificial intelligence will transform the work of lawyers*" or "*Need a lawyer? There's an algorithm for that*" are annotated as non-critical. Conversely, titles such as "*Robots put jobs at risk*" or "*Robot lawyers: how humans can fight back*" were annotated as critical. In total we annotated 6 267 article titles to determine if they contained or not a critical discourse. We checked that the coding was highly consistent between two coders.

Second, we train different standard text classification algorithms. We compare two approaches: in a first experiment, we choose to represent every title using a classic bag-of-word embedding. Each title is modeled as a binary vector, which length is equal to the vocabulary size. The vector has a non-null value at the coordinate corresponding to the words it is composed of. This vector serves as input to three classic Machine Learning models: a linear Support Vector Machine trained with Stochastic Gradient Descent, a Random Forest, and a Logistic Regression. In a second experiment, we train a sentence classification model using fastText [24].[5] fastText model architecture is inspired by the CBOW model [31], except that the middle word is replaced by the sentence label. Using n-gram features enables to use local information about the order of words in the sentence (which is useful order to keep negation structures in a critical/non-critical classification context), which is not possible with a simple bag-of-words model and is less computationally expensive than using the complete sentence with the ordering of words that is exploited by the most recent architectures such as LSTMs or Transformers.

We report evaluation metrics for our models in Table 1. As we are interested in exploiting a critical corpus containing the smallest proportion of false positives, we use the fastText model for inferring our final corpus as it has the best precision (.94) and the best F1-score (.86) to detect critical titles. Qualitatively, titles such as *"Robocops to replace British bobbies on the streets, police force reveals*", *"Growth of AI could boost cybercrime and security threats, report warns"* are classified by fastText as critical, and more ambiguous samples, as *"Google so*

[4] The train set is published on this dataverse: https://dataverse.harvard.edu/dataverse/AI_issue_mapping

[5] i.e. we do not use fastText pre-trained embedding with a subsequent algorithm, but use the complete sentence classification pipeline provided in fastText

Table 1
Performances of different classification algorithms on our human-annotated dataset. BoW stands for "Bag-of-Words"

	Critical		
Model	Precision	Recall	F1-score
BoW - Linear SVM + SGD	0.88	**0.82**	0.85
BoW - Random Forest	0.93	0.71	0.80
BoW - Logistic Regression	0.92	0.52	0.66
fastText	**0.94**	0.79	**0.86**

advanced stores will pack your products before you've thought of ordering them" are misclassified as non-critical. After running our classifier, the final proportion of articles annotated by the algorithm as critical is composed of 2 091 articles[6], accouting for 7.1% of the entire corpus.

3.3. Semantic network

We first produce the semantic network inferred from word cooccurrences observed in our sub-corpus of critical articles toward AI. To do so, we follow the methodology described in [40] which can be decomposed in three phases: term extraction, semantic similarity computation and semantic network analysis and mapping.[7] We first extracted a list of noun phrases using standard NLP tools to recognize such chunks in the full text of articles. These terms were then ranked according to their multiplicity score. The multiplicity score of a term t is inspired by the traditional GF-IDF which measures the ratio between the Global Frequency ($GF(t)$: total number of occurrences) of a term and its Document Frequency ($DF(t)$: number of distinct documents it appears in). The rationale behind such a measure is that central terms in a text are more likely to be repeated. Therefore, their GF-IDF is higher than 1. Obviously, very frequent terms will tend to repeat in a text and score high on such a metric even when irrelevant. Consequently, we use a slightly more sophisticated metric to measure the multiplicity score of a term as the ratio between the observed number of documents it appears in $DF(t)$ and the number of documents $\widehat{DF}(t)$ we should expect to observe when considering a term with the same global frequency $GF(t)$ and distributed randomly over the documents of the corpus. [8] .

$$\widehat{DF}(t) = N - N\left(\frac{N-1}{N}\right)^{GF(t)}$$

We only conserve the top 3 000 terms with the highest multiplicity score. The second step of the method consists in measuring the semantic relatedness between the terms we have short-listed. Our semantic proximity measure [40] builds on pointwise mutual information[46]. It is actually similar to the way contexts are modeled in word embedding methods such as Glove

[6] The corpus is composed of 224 articles published in 2015, 416 articles published in 2016, 389 articles published in 2017, 590 articles published in 2018 and 472 articles published in 2019

[7] Note that we used the text analysis platform CorText (https://www.cortext.net) to perform the analysis

[8] Given a term t which globally appears $GF(t)$ times in a corpus. Let's suppose its occurrences are distributed at random among the N documents that compose the corpus. Then the probability that a given document is not mentioning the term is equal to the probability that each of its occurrences fall in another document $(\frac{N-1}{N})^{GF(t)}$. The expected number of documents a randomly distributed term should not occupy is then simply obtained by summing this probability over every document in the corpus: $N(\frac{N-1}{N})^{GF(t)}$. From there it is easy to conclude that the expected number of documents a randomly distributed term should appear in is given by the equation: $\widehat{DF}(t) = N - N(\frac{N-1}{N})^{GF(t)}$

[37]. After limiting the network to pairs of terms connected with a semantic similarity above a fixed threshold of .3, we obtain a network featuring 2 991 terms (9 terms are disconnected and ignored) and 54 062 links, we identify the partition that optimizes the modularity of the network thanks to the "Louvain" algorithm [4] The final spatialization is based on a Fruchterman Reingold spatialization algorithm.[9] The node size scales with their total number of occurrences when their colors depend on the cluster the community detection algorithm has assigned them to. [10]

We think the visual depiction of the network is useful as it allows to articulate a micro-level analysis of the way words interact the one with the others, a macro reading of the structure of clusters which polarizes cluster along a line of tension (see sec 4.2) and a meso-level understanding of the way individual clusters relate the one to the others. The final visualization of the network was produced using Gephi.[11] Finally a matching algorithm assigns to each article in the corpus the cluster it is the most related to according to the overlap between its terms and cluster compositions.

4. Results

4.1. Retrieving AI technologies and application domains through graph clustering

The semantic network (see Figure 1) is computed from the enumeration of co-occurrences of terms in the article full text. The network is structured around 23 thematic clusters. Using the most central terms of each cluster we can interpret and label the topics that structure the semantic space of the corpus. The semantic clusters are populated by various types of technical entities and cover a range of application domains. The 5 largest clusters dominate the media space as they constitute 75.6% of the total number of the critical articles corpus (1 581 articles) and represent 46.4% of the network of extracted terms (1 389 terms).

The most important one, entitled "Web Algorithms" (22% of the articles in the corpus), includes articles dealing with disorders produced by web algorithms such as the ranking techniques of Facebook's newsfeed, video recommendations on Youtube or search engines such as Google. At the opposite end of the graph, the second largest cluster labelled "Future of AI" (18%), discusses the threats that the emergence of artificial intelligence and autonomous machines imitating or surpassing human capacities would be to human existence. The "Job automation" cluster (14%) contains articles warning of the risks of transformation of the labour market facing a growing robotisation process. The "Killer robots" cluster (11%) contains articles on the risks of deploying AI and autonomous machines in the context of armed conflicts. The "Facial recognition" cluster (10%) deals with various developments in facial recognition technologies in the public space, in software or on web platforms.

A second set of 9 medium-size clusters respectively represent between 1% and 4% of the total number of articles in the corpus. Together, these clusters represent 21.6% of the articles in the corpus (452 articles) and 38% of the terms in the graph (1 136 terms). They can mainly be defined by the technical entities present in the articles as "Voice Assistant", "Autonomous

[9] We tested several force directed algorithms (Fruchterman Reingold[19], Force Atlas 2[22]) and ran several iterations using various random seeding of nodes positions. The structural features of the network are so robust that the clusters systematically organize along the same axis opposing "robots" to "algorithms" (see sec. 4.2).

[10] The graph file is published on this dataverse: https://dataverse.harvard.edu/dataverse/AI_issue_mapping

[11] https://gephi.org

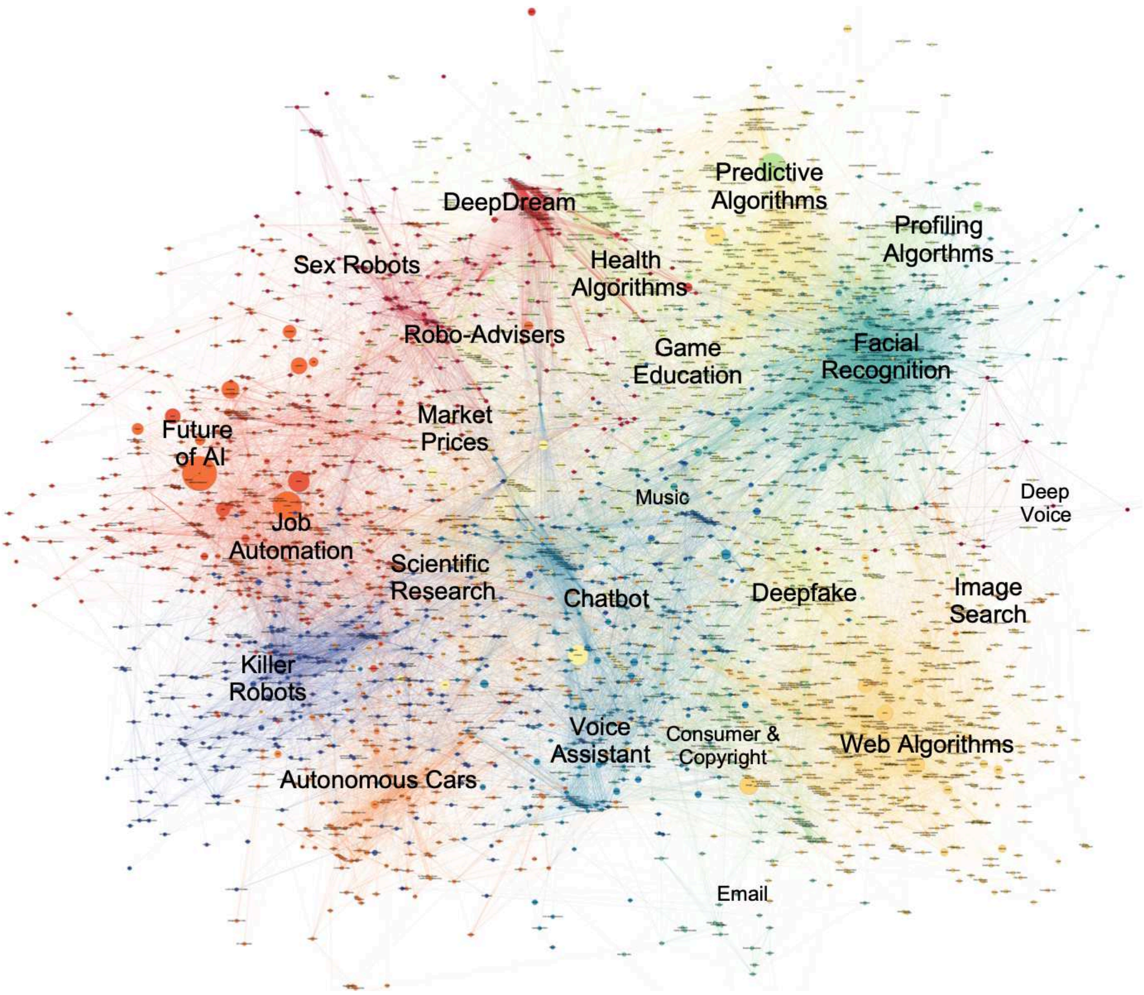

Figure 1: Our semantic network is composed of 2 991 nodes, 54 062 edges and was partitioned in 23 thematic clusters that we manually labelled. We also designed an interactive version of the network that can be consulted for exploration on this webpage

Cars", "Sex Robots", "Health Algorithms", "Deepfake", "Predictive Algorithms", "Chatbot", "Game and Education", "Profiling Algorithms".

Finally, the network is composed of 9 small clusters, each of them accounting for less than 1% of the total number of articles. These clusters focus on very specific topics, evoking types of calculators or more often fields of application such as "Robo-Advisors", "DeepDream Nightmares", "Deep Voice", "Scientific Research", "Market and Prices", "Image Search", "Email", "Music", "Consumer and Copyright".

4.2. Exploring the topological polarity between "robots" and ‘algorithms"

The network topology reveals a clear-cut separation between two main types of calculators which appears when drawing a vertical line separating the right and left side of the graph. We can thus observe a shift between articles featuring algorithmic calculation techniques incorporated into the user's environment to guide, orient or calculate his behaviors (web algorithms, facial-recognition, predictive algorithms), towards articles characterized by a personification of AI in an embodied and autonomous entity (future of AI, autonomous car, job automation, sexual and killer robots). To analyze the two poles emerging from the topological analysis, the choice was made to divide the network into two equivalent semantic spaces in terms of volume of articles and terms. To highlight the characteristics of the two semantic spaces, and following an analytical approach of sociology of translation [8], our comparative analysis focuses on three types of entities that allow us to observe the relations between technology and society: the technical entities, the human entities and the issues (see Figure 2). We devised this ontology by manual selection within the original vocabulary of the map.

The left side of the network, which we entitled “Robots”, is composed of embodied technical entities such as *robots*, *machines*, *computers*, *cars*, *weapons*, *drones*, *dolls*, etc. Other technical entities refer to very abstract and generic definitions such as *system*, *artificial intelligence*, *automation* or *model.* These devices, in addition to being physically embodied, are equipped with the ability to act autonomously without human intervention and simulate both the body and the cognitive capacities of humans. They are able to produce certain actions without human intervention in different fields such as transport (*self-driving-cars*), defense (*autonomous weapons*), labor market or physical relations (*sex dolls*). The right side of the network, which we labelled “Algorithms”, mainly concerns technical entities present in our daily digital environments. The technical entities mainly refer to precise technological devices such as *facial recognition*, *deepfakes*, *social networks*, *chatbots*, *criminal justice algorithms* and even more specifically to services embedded in our mobile terminals or on the web, sometimes associated with brands, such as *Siri*, *Search engine*, *Trending topics*, *Google Assistant*, *iphone*, *recommendation algorithm*, *Facebook Messenger*, *Google Images*, *Image search.*

The terms related to human entities reveal another difference between the two spaces mirroring each other. On the left-hand side we observe among the most frequent terms a strong presence of references to humanity, such as *human*, *humankind*, *human civilisation*, *human driver*, *human supervisor.* The notion of humanity is a generic expression that defines the entire society on which these agents pose a threat. Other terms focus on application domains in which robotic technologies are exploited, such as finance, labor market, defense and transport (*workers, customers, employees, drivers, retailers, soldiers, passengers, brokers, traders, farmers*). The space entitled ”Algorithms” is populated by more precise and personified human agents. Among the most frequent terms we identify entities referring to users of digital platforms (*Facebook users, Youtube users*), or internet accounts. The terms mainly refer to people more precisely qualified according to attributes such as their age (*children, kids, parents*), their gender (*women, men*), their ethnicity (*black people, black patients, African-Americans*), their political views (*white supremacists, black defendants, illuminati*) or finally their sexual orientation (*gays, lesbians, trans people*).

Terms related to issues are also distributed very differently across the vertical semantic frontier. On the “Robots” side, it’s the confrontation between humans and AI that is constantly called as an issue. Note the presence of terms such as *attack*, *safety*, *arms race*, *cold war*, *human extinction*, *natural disasters*, *AI-powered horror*, *mass extinction*, *physical damage*, which most often refer to war or destruction on a planetary scale, posing an existential risk to the future of humanity. These threats give rise to issues of control of these autonomous technologies, such as *ban*, *petition*, *human oversight*, *lack of accountability*, *super-intelligence control problem*, *control problem.* In the space entitled “Algorithms” other forms of critical discourse emerge that refer to legal issues. Indeed, we find a semantic field made up of legal references such as *crime*, *law enforcement*, *Human Rights*, *lawsuit*, *Civil Liberties*, *prejudice*, *fraud*, *public interest.* The disorders produced by technical agents denounced in the articles concern discrimination

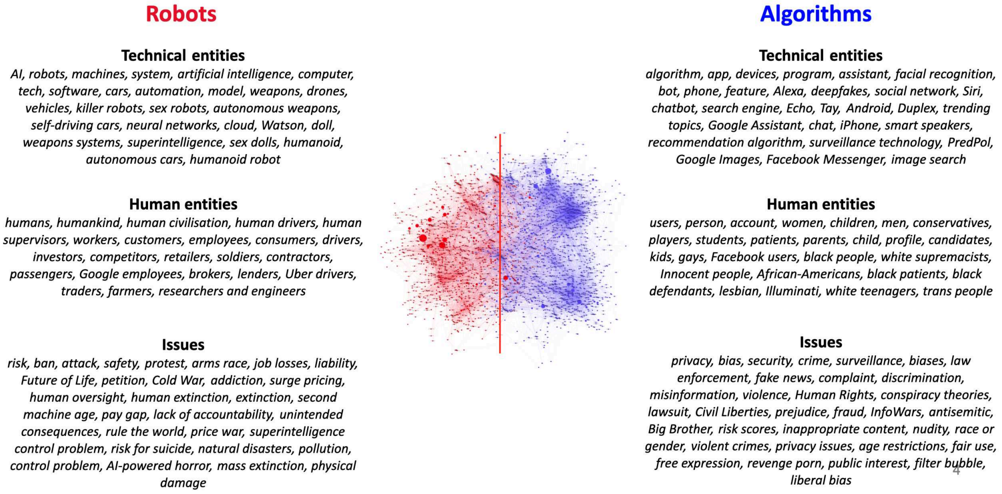
Robots
Technical entities
AI, robots, machines, system, artificial intelligence, computer, tech, software, cars, automation, model, weapons, drones, vehicles, killer robots, sex robots, autonomous weapons, self-driving cars, neural networks, cloud, Watson, doll, weapons systems, superintelligence, sex dolls, humanoid, autonomous cars, humanoid robot
Human entities
humans, humankind, human civilisation, human drivers, human supervisors, workers, customers, employees, consumers, drivers, investors, competitors, retailers, soldiers, contractors, passengers, Google employees, brokers, lenders, Uber drivers, traders, farmers, researchers and engineers
Issues
risk, ban, attack, safety, protest, arms race, job losses, liability, Future of Life, petition, Cold War, addiction, surge pricing, human oversight, human extinction, extinction, second machine age, pay gap, lack of accountability, unintended consequences, rule the world, price war, superintelligence control problem, risk for suicide, natural disasters, pollution, control problem, AI-powered horror, mass extinction, physical damage
Algorithms
Technical entities
algorithm, app, devices, program, assistant, facial recognition, bot, phone, feature, Alexa, deepfakes, social network, Siri, chatbot, search engine, Echo, Tay, Android, Duplex, trending topics, Google Assistant, chat, iPhone, smart speakers, recommendation algorithm, surveillance technology, PredPol, Google Images, Facebook Messenger, image search
Human entities
users, person, account, women, children, men, conservatives, players, students, patients, parents, child, profile, candidates, kids, gays, Facebook users, black people, white supremacists, Innocent people, African-Americans, black patients, black defendants, lesbian, Illuminati, white teenagers, trans people
Issues
privacy, bias, security, crime, surveillance, biases, law enforcement, fake news, complaint, discrimination, misinformation, violence, Human Rights, conspiracy theories, lawsuit, Civil Liberties, prejudice, fraud, InfoWars, antisemitic, Big Brother, risk scores, inappropriate content, nudity, race or gender, violent crimes, privacy issues, age restrictions, fair use, free expression, revenge porn, public interest, filter bubble, liberal bias
4

Figure 2: Examples of higher-occurrence terms extracted from the two parts of the network (Robots and Algorithms) related to technical entities, human entities and issues

(*bias, biases, discrimination, antisemitic, race or gender, fair use, risk score, liberal bias*), privacy issues (*privacy, surveillance, Big Brother, privacy issues*), the difficulties of filtering or exposure to inappropriate content (*violence, inappropriate content, nudity, age restriction, violent crime*) fake news (*fake news, misinformation, conspiracy theories, revenge porn, filter bubble*), or censorship and freedom of expression (*free expression*).

4.3. Analyzing issue-related verbs and temporality markers

The graph topology thus shows an opposition between two distinct subsets which are characterized by different technical and human entities but also different issues. In order to further analyze the differences in the way AI is criticized in the two semantic spaces we opted for partitioning our corpus of articles in two parts corresponding to the "Robots" and "Algorithms" perspectives We split the set of clusters in the two following groups:

- **Algos**: "Web Algorithms", "Facial Recognition", "Voice Assistant", "Consumer & Copyright", "Email", "Music", "Deep Voice", "Image Search", "Profiling Algorithms", "Game & Education","Chatbot","Predictive Algorithms", "Deepfake","Health Algorithms"
- **Robots**: "DeepDream Nightmare", "Market & Prices", "Scientific Research", "Robo-Advisors", "Sex Robots", "Autonomous Cars", "Killer Robots", "Job Automation", "Future of AI"

This split is visualized figure 2. The 9 clusters composing the Robots meta-cluster is rich of 1 094 articles. There are 14 clusters contributing to the "Algorithmic" side which is populated by 997 articles.

We then proceed to a comparative analysis of the relative frequency of issue-related verbs and time markers used in both sub-corpora:

- **Issue-related verbs.** They were built from a first naive extraction of the most frequent verbs.[12] The list was then manually curated to extract a set of 66 verbs that relate to

[12] We extracted the 1000 most frequent verbs using the same procedure as for terms (see subsection 3.3

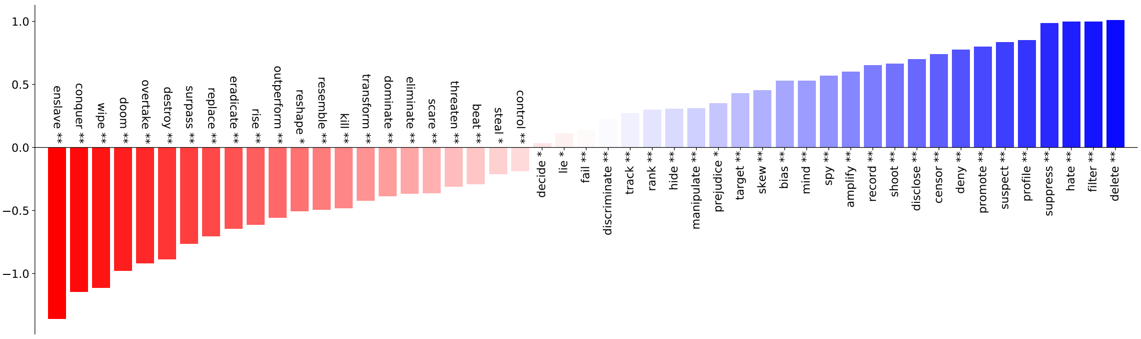

Figure 3: s coefficients for the 49 issue-related verbs. Verbs which are over-represented in the "Robots" subcorpus show in red. The number of stars after each entity relate to the p-value of the associated Fisher exact test (one star if $p < .05$, two stars if $p < .01$)

issues produced by the computing entities.

- **Time markers.** We ran the Named Entity Recognizer from Spacy [21] to identify the 1000 most frequent temporal entities. Again, this list was manually curated in order to keep entities referring without ambiguity to a temporal dimension.[13] Our final list of temporal markers is rich of 109 entities.

We then measure the relative frequency of temporal markers and issue-related verbs in both sub-corpora. We run a Fisher exact test to appreciate whether the frequency of each entity is over-represented in one of the two corpora. 49 verbs and 35 time markers pass the test (with a p-value $p < .05$). Figure 3 and Figure 4 plot the ratio of entities' frequencies between the Algorithm and Robot subcorpora, showing which entities are particularly over/under-represented on both side of our network. More precisely, for an entity i, we measure and plot the following score: $s(i) = log(\frac{p(i|Algos)}{p(i|Robots)})$. $s(i)$ is positive when the relative frequency of the entity i is used more often in the Algorithms subcorpus. Conversely, negative values correspond to entities which are concentrated in the Robots subcorpus. The height of bars measures how important the deviation is and the statistical test allows to check that such over/under-representation is indeed statistically significant.

The verbs relating to the most important issues in the subcorpus "Robots" express a threat from machines and autonomous intelligences belonging to a prophetic, even apocalyptic discourse. We find verbs that express notions of destruction (*doom, destroy, eradicate, kill, eliminate*), domination of machines (*enslave, dominate, conquer*), but also of overtaking or replacing humans by these technical entities (*overtake, surpass, replace*). Other verbs refer to notions of transformation and change (*reshape, transform*) or to the capacities of these technical agents to imitate or simulate human behaviors (*resemble, simulate*). In the articles associated with 'Algorithms" the verbs refer to issues of filtering and censoring information (*filter, delete, suppress, censor*), surveillance and privacy issues (*profile, suspect, spy, target, track*), denouncing forms of discrimination (*bias, discriminate*) or the propagation and amplification of fraudulent content (*promote, amplify*).

It is interesting to understand how temporality is expressed within critical discourse. Research on argumentative forms in the sociology of controversy [10, 11] invites us to look at

[13] Contextual entities, referring to temporal markers in a vague way (e.g. "several years", "winter"), frequencies (e.g. "everyday", "weekly"), specific events (e.g. "Christmas") were removed.

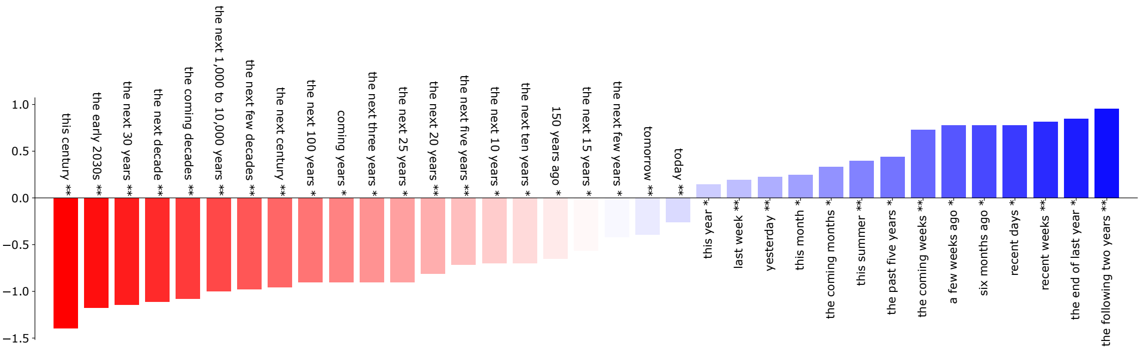

Figure 4: s coefficients for the 35 temporal entities. Temporal entities which are over-represented in the "Robots" subcorpus show in red.The number of stars after each entity relate to the p-value of the associated Fisher exact test (one star if $p < .05$, two stars if $p < .01$).

the way in which temporal scales and the associated regimes of enunciation of the future are deployed. When using temporal markers, actors offer clues about the actual temporality of the problem at stake: is the threat immediate, or long term?

On the side composed of the articles associated with the semantic space "Robots", the named entities related to time markers are associated with a way longer term temporality scale, whether they refer to the past or the future. The majority of markers are structured around expressions about the future starting with the expression "the next" and associated with temporalities often counted in tens or hundreds of years (*the next 10 years, the next 30 years, the next 100 years, the next 1,000 to 10,000 years*). Other markers have the particularity of not precisely defining the timeframes to which they refer (*the next decade, the next century, the next few decades*). Other markers do not refer to the future but to the past and are characterized by the fact that they also refer to a distant projection into the past (*150 years ago*). In the articles associated with "Algorithms", the extracted named entities related to time mostly refer to the present or a closer past often expressed in days, weeks or months (*recent days, recent weeks, six months ago*), but also to a very close future in comparison with the other part of the corpus (*the coming weeks, the coming months, the following two years*).

The analysis of the issue-related verbs and the temporal named entities confirms our first analyses of a polarization in the corpus of press articles. On the one hand, critical articles that develop arguments on a distant and threatening future of robots and autonomous AI that need to be controlled. On the other hand, we find arguments advocating for a regulation of technologies based on algorithms used in our daily life, in order to limit the problems of discrimination, privacy or content filtering.

5. Conclusion

Using NLP methods, we have identified a corpus of press articles producing a critical discourse on AI and associated technologies. By conducting a semantic network analysis, we explored the technical and human entities that populate the two topological poles that emerge from the structure of our semantic network. We also analyzed verbs and time related named entities to analyze which issues and temporalities the critical discourse on AI is composed of.

We have identified two opposing views in the semantic space that seem to coexist in the media space. This dual perspective is reminiscent of the long history of the relationship between computer technology and society. Classically one distinguishes between, on the one hand, the scientific project to develop the intelligence of machines with cognitive capacities with the aim of reproducing human reasoning, and on the other hand, the project to increase the intelligence of humans thanks to computer technologies and to equip the human environment with communication and calculation tools [30].

We have also demonstrated that these two types of technologies are associated with two different regimes of criticism. The first expresses fear of autonomous technologies that focus on their ability to simulate, surpass, replace or exterminate humanity and represent a threat that needs to be controlled. This regime of criticism on the fear of robots and autonomous AI is fueled by popular representations of these technologies coming from science fiction and turning intelligent machines into mythical creatures [34]. This regime is also associated with a religious discourse which contains a prophetic dimension on the future of humanity [43, 42]. The other regime of criticism concerns the technologies that compose our daily digital environments. It is more rooted in a discourse of social criticism and injustices (censorship, discrimination, surveillance) towards specific populations, and focus on regulatory issues concerning the way in which the propagation of information is managed (exposure, amplification).

Acknowledgments

This research received funding from Good In Tech Chair, under the Fondation du Risque in partnership with Institut Mines-Télécom and Sciences Po. This research was partly supported by the GOPI project (ANR-19-CE38-0006) and by the French government under management of Agence Nationale de la Recherche as part of the ”Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

[1] P. Barberá, A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker. “Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data”. In: *American Political Science Review* 113.4 (2019), pp. 883–901.

[2] Y. Benkler, R. Faris, and H. Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics.* Oxford University Press, 2018.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent dirichlet allocation”. In: *the Journal of Machine Learning research* 3 (2003), pp. 993–1022.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[5] L. Boltanski. “Sociologie critique et sociologie de la critique”. In: *Politix. Revue des sciences sociales du politique* 3.10 (1990), pp. 124–134.

[6] J. Brennen. “An industry-led debate: How UK media cover artificial intelligence”. In: (2018).

[7] T. Bucher. “The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms”. In: *Information, communication & society* 20.1 (2017), pp. 30–44.

[8] M. Callon. “Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay”. In: *The sociological review* 32.1_suppl (1984), pp. 196–233.

[9] D. Cardon and M. Crépel. “Algorithmes et régulation des territoires”. In: *Gouverner la ville numérique, La vie des idées* (2019), pp. 83–102.

[10] F. Chateauraynaud. “Regard analytique sur l’activité visionnaire”. In: *Du risque à la menace. Penser la catastrophe, Paris, PUF* (2013), pp. 309–389.

[11] F. Chateauraynaud and J. Debaz. “Agir avant et après la fin du monde, dans l’infinité des milieux en interaction”. In: *Multitudes* 3 (2019), pp. 126–132.

[12] C.-H. Chuan, W.-H. S. Tsai, and S. Y. Cho. “Framing Artificial Intelligence in American Newspapers”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 339–344.

[13] J.-P. Cointet and S. Parasie. “What Big data does to the sociological analysis of texts? A review of recent research”. In: *Revue francaise de sociologie* 59.3 (2018), pp. 533–557.

[14] A. Edelman, T. Wolff, D. Montagne, and C. A. Bail. “Computational Social Science”. In: *Annual Review of Sociology* 46 (2020).

[15] J. A. Evans and P. Aceves. “Machine translation: Mining text for social theory”. In: *Annual Review of Sociology* 42 (2016), pp. 21–50.

[16] E. Fast and E. Horvitz. “Long-term trends in the public perception of artificial intelligence”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[17] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar. “Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI”. In: *Berkman Klein Center Research Publication* 2020-1 (2020).

[18] L. Floridi. “Translating principles into practices of digital ethics: Five risks of being unethical”. In: *Philosophy & Technology* 32.2 (2019), pp. 185–193.

[19] T. M. Fruchterman and E. M. Reingold. “Graph drawing by force-directed placement”. In: *Software: Practice and experience* 21.11 (1991), pp. 1129–1164.

[20] M. Gentzkow, B. Kelly, and M. Taddy. “Text as data”. In: *Journal of Economic Literature* 57.3 (2019), pp. 535–74.

[21] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: 10.5281/zenodo.1212303. URL: https://doi.org/10.5281/zenodo.1212303.

[22] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. “ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software”. In: *PloS one* 9.6 (2014), e98679.

[23] A. Jobin, M. Ienca, and E. Vayena. “The global landscape of AI ethics guidelines”. In: *Nature Machine Intelligence* 1.9 (2019), pp. 389–399.

[24] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. “Bag of Tricks for Efficient Text Classification”. In: *Eacl*. 2017.

[25] B. Latour. *Changer de société, refaire de la sociologie*. La découverte, 2014.

[26] P. F. Lazarsfeld and R. K. Merton. *Mass communication, popular taste and organized social action*. Bobbs-Merrill, College Division, 1948.

[27] P. F. Lazarsfeld and A. R. Oberschall. “Max Weber and empirical social research”. In: *American sociological review* (1965), pp. 185–199.

[28] B. Liu. “Sentiment analysis and opinion mining”. In: *Synthesis lectures on human language technologies* 5.1 (2012), pp. 1–167.

[29] A. Machut. “Julia Cagé, Nicolas Hervé, Marie-Luce Viaud, L’information à tout prix, Ina Editions, 2017.” In: *Revue française de sociologie* (2018).

[30] J. Markoff. “Machines of loving grace: The quest for common ground between humans and robots”. In: Ecco New York. 2015.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL].

[32] B. Mittelstadt. “Principles alone cannot guarantee ethical AI”. In: *Nature Machine Intelligence* 1.11 (2019), pp. 501–507.

[33] J. W. Mohr, R. Wagner-Pacifici, R. L. Breiger, and P. Bogdanov. “Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics”. In: *Poetics* 41.6 (2013), pp. 670–700.

[34] S. Natale and A. Ballatore. “Imagining the thinking machine: Technological myths and the rise of artificial intelligence”. In: *Convergence* 26.1 (2020), pp. 3–18.

[35] C. O’neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

[36] F. Pasquale. *The black box society*. Harvard University Press, 2015.

[37] J. Pennington, R. Socher, and C. D. Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[38] R. Perrault, Y. Shoham, E. Brynjolfsson, J. Clark, J. Etchemendy, B. Grosz, T. Lyons, J. Manyika, S. Mishra, and J. C. Niebles. “The AI index 2019 annual report”. In: *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA* (2019).

[39] A. Rosenblat. *Uberland: How algorithms are rewriting the rules of work*. Univ of California Press, 2018.

[40] A. Rule, J.-P. Cointet, and P. S. Bearman. “Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014”. In: *Proceedings of the National Academy of Sciences* 112.35 (2015), pp. 10837–10844.

[41] N. Seaver. “Algorithms as culture: Some tactics for the ethnography of algorithmic systems”. In: *Big Data & Society* 4.2 (2017), p. 2053951717738104.

[42] B. Singler. ““Blessed by the Algorithm”: Theistic Conceptions of Artificial Intelligence in Online Discourse”. In: *AI & society* 35.4 (2020), pp. 945–955.

[43] B. Singler. “Existential Hope and Existential Despair in Ai Apocalypticism and Transhumanism”. In: *Zygon* 54.1 (2019), pp. 156–176.

[44] O. Stuhler. “What’s in a category? A new approach to Discourse Role Analysis”. In: *Poetics* (2021), p. 101568.

[45] M. Vergeer. “Artificial intelligence in the dutch press: An analysis of topics and trends”. In: *Communication Studies* 71.3 (2020), pp. 373–392.

[46] J. Weeds, D. Weir, and D. McCarthy. “Characterising measures of lexical distributional similarity”. In: *COLING 2004: Proceedings of the 20th international conference on Computational Linguistics*. 2004, pp. 1015–1021.