Knowledge Technologies for Democracy

Project Number: 101094302 Start Date of Project: 01/02/2023 Duration: 36 months

Deliverable 3.1

Social Risk Toolkit - Modules A, B and C

Dissemination Level	Public (PU)
Due Date of Deliverable	31/01/2024
Actual Submission Date	26/01/2024
Work Package	WP3 Advancing the state of the art: Civic Participation and Knowledge Technologies
Task	T3.1 Establishing a conceptual overview of the problems AI and big data pose for democracy: perspectives of individuals
Type	Report
Version	V0.1
Number of Pages	p.1 – p.56

Deliverable Abstract

This document presents the theoretical framework guiding the first three modules of the KT4D Social Toolkit. From its second iteration, it presents the progress made beyond the state of the art, and the methodology adopted to provide new data and an original theoretical framework, as the project progresses.

Funded by the European Union

KT4D has received funding from the EU's Horizon Europe research and innovation programme under Grant Agreement no. 101094302.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

DELIVERY SLIP

Date	Name	Partner/Activity	Date
From:			
Moderated by:	Tiffany Morisseau	STRANE	15/01/2024
Reviewed by:	Lucia Garcia	CIB	24/01/2024
Approved by:	Eva Power	TCD	26/01/2024

DOCUMENT LOG

Issue	Date	Comment	Author	ORCID ID
V0.1	15/01/2024	First draft	Tiffany Morisseau Eleonora Lima	0000-0002-2523-7755 0000-0001-7578-8005
V0.2	19/02/2024	Peer Review	Eva Power	0009-0000-5118-0417
V0.3	20/01/2024	Addressing peer review comments	Tiffany Morisseau Eleonora Lima	0000-0002-2523-7755 0000-0001-7578-8005
V0.4	24/01/2024	Peer Review	Lucia Garcia	0009-0006-4257-7993

TERMINOLOGY

Terminology/Acronym	Definition
AI	Artificial Intelligence
API	Application Programming Interface
DSA	Digital Services Act
GDPR	General Data Protection Regulation
GLM	General Language Model

Funded by the European Union

KT4D has received funding from the EU's Horizon Europe research and innovation programme under Grant Agreement no. 101094302.

Table of Contents

Executive Summary

The integration of artificial intelligence in multiple aspects of our daily lives poses a great challenge in shaping civic participation within societies. Effective civic engagement finds its foundation in the empowerment of citizens who actively participate in public discourse, and in the articulation of their viewpoints. Understanding the dynamics that underpin trust in democratic institutions in this new context is thus important.

The aim of the first three modules of KT4D's Social Risk Toolkit thus focuses on the individual aspects of this challenge and is multifaceted. Module A questions individuals' capacity to navigate AI-based content and recommendations while safeguarding their autonomy and free will. This module serves as a guide to enhance people's ability to assess the quality and diversity of information they encounter, and to determine where exactly the autonomy and free will to which anyone can rightfully aspire is threatened.

Module B examines the requisites for the utilisation of AI to preserve institutional trust. Indeed, the use of AI in a wide range of fields could generate situations that are, or are perceived to be, unfair, thereby threatening the social balance established between members of the same community. A first important +step is thus to understand what is really a source of concern, based on an in-depth examination of the situations likely to alter this trust, before crafting regulations to remedy these situations. By exploring diverse scenarios, the experimental component of this module will prompt individuals to introspect on their values concerning human productions, such as creativity, equity, and effort.

Module C puts the new challenges posed by technological transformations into perspective, by analysing historical precedents and how they have shaped culture and social interactions over centuries.

These three modules therefore look into people's perspective on the potentials and pitfalls associated with AI and big data, by the means of a socio-cognitive and historical perspective. Through a comprehensive exploration and assessment, this part of the toolkit will seek to address an important challenge: identify exactly where regulation is needed to ensure that AI benefits society as a whole, and identify the conditions for trust in social institutions, in order to guarantee effective public debate and societal choices are made by citizens themselves. The historical perspective developed in Module C will allow us to put into perspective the concerns identified by both individuals and experts on AI.

1 Module A: AI, free will and autonomy

1.1 Introduction

In this first module, we explore the ways in which artificial intelligence can affect individuals' free will and autonomy.

According to *The Stanford Encyclopedia of Philosophy*:

> The term "free will" has emerged over the past two millennia as the canonical designator for a significant kind of control over one's actions [with] Questions concerning the nature and existence of this kind of control (e.g., does it require and do we have the freedom to do otherwise or the power of self-determination?), and what its true significance is (is it necessary for moral responsibility or human dignity?).

Free will is thus the ability of individuals to determine freely and on their own how to act and think. In the context of artificial intelligence, the philosophical question of free will arises as more and more choices, decisions, opinions, and attitudes seem to be constrained by the way the digital world is structured. In particular, the question of the information that determines our choices is at the heart of the matter: how can we be free as individuals, and as members of a society, if we are unable to assess the value of the information to which we are exposed?

The first part of module A therefore explores the central question of the relationship to information in a world where AI is used on a massive scale. In particular, we will look at the extent to which individual opinions are influenced by algorithms proposed by private parties, with their own motivations. We will distinguish between the knowledge required for informed decision-making at the individual level, and collective decision-making. To ensure that citizens' perspectives are truly incorporated in political decision-making, individuals must indeed not only possess the capacity but also the motivation to cultivate well-informed opinions. The propensity to form opinions that accurately reflect reality, underpinned by evidence-based reasoning, is at the heart of effective collective decision-making processes.

Regarding the *ability* to form informed opinions, we will look into the current scientific knowledge on how misinformation actually influences people's representations of the world. Regarding the *motivation* to form (and share) sound opinions that are necessary to make efficient collective decisions, we will clarify how information shared on social media, with or without malicious intent, can play a major role in altering the quality of public debate. Not necessarily because people incorporate false information into their belief system, but because it may contribute to creating a climate of mistrust in institutions. The massive use by social networks of algorithms, designed to optimise user engagement, appears to be amplifying this phenomenon. How access to quality information has been undermined by the massive use of AI and algorithms in digital media, and how these shape individuals' online behaviours, reasoning faculties and opinions, remains extremely important questions, whose answers are not quite settled yet.

The second part of this module will focus more directly on the issue of individual autonomy and self-governance. That is, individuals' ability to effectively make their own decisions and act independently, and to determine their own rules and objectives. In ethics and bioethics, patient autonomy is a key principle that emphasises a person's right to make decisions about their own medical treatment. Many of our online experiences suggest that individual autonomy cannot be taken for granted. New forms of subordination have emerged that do not arise from censorship or physical constraint, but from a voluntary submission to our immediate desires, and from the way the choices that we come to make online are presented to us. We are asked to accept terms and conditions, but to what extent do we truly accept these terms, if we are not in the position to do otherwise?

In this Deliverable, we will set out the main conclusions drawn from the literature in psychology on these issues. The next version of this document will propose a series of experimental protocols aimed at clarifying the boundaries of informed use of AI-based algorithms and systems.

1.2 State of the art: literature review

1.2.1 The problem of information quality

This first section draws on the existing literature on the psychology of belief, and the work that has been carried out in recent years on the impact of social networks and misinformation on citizens' opinions and attitudes. It explores the problem of access to high-quality information, focusing on two critical aspects: the ability to differentiate between accurate and inaccurate information, essential for collective decision-making, and the importance of accessing a wide range of information sources. From this starting point, we will identify where regulations and recommendations should focus.

1.2.1.1 What makes information attractive?

The interplay between the online information ecosystem and human cognitive processes has been extensively studied from a variety of perspectives. Over the last decade, there has been a great deal of literature on the ease with which false information spreads, with some showing, for example, that false information spreads six times faster than true information. (Vesoughi et al., 2018[1]).

Fake news are all the more successful as they are detached from reality, and crafted to possess a cognitive appeal that genuine news struggles to match. These fabricated narratives frequently include elements of heightened social relevance and implications, thereby enhancing their cognitive attractiveness (Acerbi, 2019). They can be manufactured based on features that make them attractive in an almost unconstrained way, whereas “true” news cannot, simply because they need to correspond to reality. False information can be intentionally crafted to propagate more extensively than accurate information. Shallow engagement will favour content that elicits quick reactions, in the form of likes and comments, as opposed to content that is in-depth, truthful, or relevant.

Some information also holds an inherent appeal due to its contribution to an influential explanatory framework, which, if validated, offers a coherent understanding of reality. Comprehending reality is a catalyst for action, and generally, individuals tend to favour information that possesses explanatory implications concerning elusive aspects of the world (Lantian et al., 2020). Moreover, research suggests a propensity for individuals to gravitate toward conspiracy theories when attempting to rationalise events that appear inexplicable to them (Lantian et al., 2020). Furthermore, there exists a tendency for individuals to find conspiracy theories more enticing, particularly when experiencing frustration in essential psychological

[1] This famous study by Vosoughi et al. showed that a small number of fake news stories can be highly influential and reach between 1,000 and 100,000 people. Yet, these figures need to be balanced against the potential dissemination of true information in general (not just of rumours that have been proven to be true) and the latter is, in fact, much more widely and rapidly disseminated (Acerbi, 2019).

needs, such as social exclusion (Graeupner & Coman, 2017). This intrinsic quest for meaning often leads individuals to prioritise straightforward information that offers a robust explanatory framework, potentially overshadowing more intricate yet factually accurate information that may hold less personal relevance.

1.2.1.2 To what extent do people believe in fake news and deepfakes?

For many scholars[2], the rise of digital communications and in particular social media is one of the main, if not the only, cause of a so-called "infodemia". However, it seems that the effect of social networks on the toxic nature of false information is actually elsewhere.

That lies spread faster than truths is not a recent concern. More than three centuries ago, Jonathan Swift lamented that "if a lie be believed only for an hour, it hath done its work, and there is no further occasion for it. Falsehood flies, and truth comes limping after it." Misinformation is not a social media phenomenon. It was present before the digital age (Altay, Berriche & Acerbi, 2023), and one can easily find examples of misinformation disseminated by traditional media such as newspapers and television.

The study of misinformation should be contextualised within the entire information ecosystem (Altay, Berriche & Acerbi, 2023). People are much more likely to be exposed to information from a variety of sources that disseminate ideas in an approximate way. While some online players do indeed aim for outright disinformation (Elswah & Howard, 2020), most people who share information on the Internet do not do so with the malicious intention of misleading others. Conveying genuine information is not always the primary goal of communication, especially when it comes to communication on the internet (Acerbi, 2019), and content shared on social media does not carry a strong presumption of truthfulness. Several results support this interpretation. First, it seems that people are not interested in the truthfulness of what they share. Moreover, people commonly share information without even reading it (Gabielkov et al. 2016). Research on persuasion suggests that individuals are not so easily influenced by the contents they encounter, but rather the reverse is true: individuals will seek out information that suits them and their needs.

When it comes to issues where truthfulness does matter from a person's point of view, there are good reasons to believe that they are more vigilant. Certainly, information circulates on the internet with varying degrees of truthfulness and manipulative intentions, and it is often crafted to be cognitively attractive and more responsive to our expectations. But our ability to monitor the intentions of others works both ways, and we are not so easily fooled. Fake news represents a very small percentage of the information consumed in the US (0.15% according to Allen et al., 2021) and 80% of total exposure to fake news can be attributed to a very small number (1%) of individuals (Grinberg et al., 2019).

Deepfakes seem to constitute a greater challenge. Deepfakes are videos that are manipulated using artificial intelligence to appear as though someone is saying or doing something they did not. Some studies (e.g., Dobber et al., 2021) suggest that they can significantly influence political attitudes, especially when they are micro-targeted to specific demographic groups. Deepfakes would thus be a powerful mode of disinformation, with other forms such as false news stories or trolling on social media platforms. For example, a deepfake video discrediting a political candidate can negatively affect people's attitudes towards both the politician and their associated party. Yet, other studies show that deepfakes are more effective when they are shown to demographic groups that are already personally interested in their content. Strong supporters of a

[2] "Post-truth" became the Oxford Dictionaries' Word of the Year in 2016.

candidate will disregard a deepfake video discrediting their candidate, due to their existing beliefs and loyalties.

In actual fact, deepfakes are revolutionary, but as a continuation of existing manipulation tactics. They are one more method in a long history of information manipulation that includes rhetoric, propaganda, and deceptive presentation of facts (Etienne, 2021). The narrative that deepfakes pose an unprecedented threat to information trustworthiness is not as solid as it seems. While deepfake technology is rapidly advancing, it is neither the first nor the most powerful tool for manipulating information. Deepfakes themselves are not the cause of a loss of trust among the public: distrust in political representatives and institutions predate these technologies.

Instead of deepfakes contributing to a post-trust era, they might intensify existing scepticism or the erosion of trust, leveraging generalised distrust that already exists in society. Deepfakes could actually aid in the transition from instrumental rationality to social rationality by fostering a critical approach to information consumption and trust-building in the digital age. Just as societies learned to question the authenticity of photographs, the presence of deepfakes may teach people to develop a more critical and informed approach to digital content, encouraging the search for trustworthy sources of information.

The case of deepfake videos is enlightening in that it reveals the point at which belief in the veracity of information communicated can actually pose a problem. If someone is interested in a particular issue (because there is a great deal at stake for the individual), then they will be more wary.

1.2.1.3 Do algorithms influence people's opinions?

Contrary to the belief held by many political consultants and scholars, the political influence of bots is in fact often exaggerated. Most forms of political persuasion, including online advertising and content promoted by bots, have little effect on changing peoples' votes (Broockman & Green, 2014; Nyhan, 2018). Persuading people to change their votes, especially committed partisans, is quite difficult. Existing research indicates that various forms of campaign actions, including TV advertisements and online ads, have very small effects or no effect at all. The effects are likely to be even smaller with fake news and bot-promoted content due to the polarisation in politics. Dubious political content online usually reaches already-informed people with strong opinions, who are less likely to change their perspectives due to such content. Finally, statistics about the reach of fake news or bot activities are often not that reflective of real-world scenarios, because a significant proportion of the bot followers might also be bots, or and because the shared content might not reach many people. In sum, while bots are concerning due to their potential to mislead and polarise citizens, undermine trust in media and distort public debate, their impact on election outcomes is likely limited.

For instance, Kalla & Broockman (2018) conducted a series of meta-analyses of field experiments and quasi-experiments showing that campaign persuasion in general elections is relatively minimal. Persuasion in such environments could occur under specific circumstances, but these effects typically decay rapidly or are likely to be negligible closer to election day. Voters are likely to have been repeatedly exposed to campaign-related information and tend to retain as much information as they care to hold onto. The effect of further campaign efforts, therefore, becomes marginal. Campaign attempts to make certain considerations more salient are likely to have limited effects in general elections, especially when individuals are exposed to competing arguments and information.

In summary, while online advertising is an increasingly popular method for mass communication, particularly in political campaigns, its effects have rarely been evaluated systematically, and as a consequence, the ability of political ads to influence voting behaviour or preferences remains, at the very least, very disputed. Some users may be more likely to recall the ads, but there is no significant effect on their recognition or positive evaluation of the candidates depicted in the ads. This suggests that even frequent exposure to political messages online may be insufficient to convey new information or change attitudes anyways.

1.2.1.4 Algorithms, toxic content and polarisation

From an individual's perspective, the utility of knowing the truthfulness of a particular fact is limited and depends on the context. The importance of an accurate representation of reality depends on the intended use of the information – it may be crucial or just an added benefit. Truthfulness is not always a prerequisite for endorsement or sharing, as the value of information often extends beyond its accuracy. Plausibility, rather than certainty, can suffice based on the purpose at hand, whether for oneself or in communication with others. One might fervently support information solely because it seems plausible, without looking too deep into its actual likelihood of being true. Endorsing and sharing information are interconnected, as people generally share what they endorse to some extent, yet they may also find value in others' entertaining diverse beliefs. While completely false information is typically avoided, uncertain details can be socially relevant and easily shared if the potential reputational consequences, in case of falsehood, are deemed acceptable.

From this perspective, algorithms can very much contribute to distorting the logic of informational exchanges on the Internet, in particular by encouraging the polarisation of exchanges, because the aim of social media platforms is to increase user engagement. Understanding the intricate and often opaque nature of interactions between users and the algorithms that curate content on social media and other online platforms is crucial but challenging due to limited transparency and the dynamic nature of the algorithms (Lewandowsky, Robertson and DiResta, 2023). How social media algorithms work remains very opaque, but timelines will certainly favour content that has been the subject of higher engagement. An obvious consequence is that posts prompting for an immediate reaction will climb to the top of the newsfeed. Pictures of a cat doing a funny face will have no real consequences, but the same is true for controversial political opinions, which will cause many comments, or for sensational stories that will generate more interest and will be "liked" more.

The truthfulness of information is less relevant when the paramount aspect lies in the ability of the induced emotion to be justified and shared among individuals (Schaffner & Luks, 2018). Substantial experimental evidence corroborates the widespread phenomenon of emotional transmission across social networks (Kramer, Guillory & Hancock, 2014; Martel, Pennycook & Rand, 2020). Emotions serve as influential drivers for dissemination within social media spaces, particularly when expressing negative sentiments or moral emotions (Brady et al., 2017). Recent research conducted by Rathje, Van Bavel & van der Linden (2021), analysing 2.7 million posts on Facebook and Twitter, underscores the strong predictive nature of content targeting out-groups in driving engagement across social media platforms. The quality of public debate thus seems to be threatened by increased polarisation: recommendation engines, with the aim of keeping users for as long as possible, come up with increasingly polarising content. Such dynamics can even lead to violent behaviours, as shown by the role of social media in organising the storming of the U.S. Capitol on January 6, 2021.

A model proposed by Santos, Lelkes and Levin (2021) shows how the strength of social influence affects opinion dynamics, by stressing the role of link recommendation algorithms commonly used in online social networks in the creation of ties and opinion polarisation. Algorithms tend to suggest new connections between users with a high number of common acquaintances, which can promote structural similarity and the formation of isolated, like-minded communities. In this context, even moderate opinions can contribute significantly to opinion polarisation. Other empirical data suggests that link recommendation algorithms do indeed change the rewiring pattern of social networks.

Moreover, attitudinal disagreements are often coupled with mistrust and contempt (reflected by the notion of affective – as opposed to attitudinal – polarisation) and these can be fed by moral and emotional content (Van Bavel & Pereira, 2018). Groups with different interests within polarities can also converge and fuel a movement of global polarisation (Benkler, Faris & Roberts, 2018).

1.2.1.5 Implication in terms of citizen's ability to contribute to collective decision-making

The consequences of information are not the same at the individual and societal levels. Upholding distorted depictions on certain topics, especially those with significant political and social ramifications, can result in profound effects, impacting areas such as health and politics. The success of political endeavours relies on their adherence to a truthful representation of reality. The collective understanding of this representation holds significant consequences when influencing the outcomes of decisions made by society as a whole. Democracy indeed requires a common base of knowledge among citizens to function optimally, which includes trust in electoral processes and evidence to inform policy debates (Lewandowsky et al., 2023).

Research shows that widespread disinformation campaigns are eroding the shared knowledge that democracy depends on, thus undermining democratic institutions and processes. Not because people believe in this false information, but because they have less trust in the benevolence of media institutions. Even when the epistemic commitment to misconceptions about the world is not very strong and may only be a means of justifying pre-existing intuitions, individual attitudes can be harmful when they become dominant in the public sphere. Once a narrative enters public discourse, it has the potential to evolve into a norm that shapes the perception of societal issues. Consequently, modes of thinking that prioritise coalition-building over collaboration and truth-seeking may dominate more moderated forms of discourse. Returning to a peaceful debate is then very difficult, as people are even less interested in hearing contradictory views. Positions defended for social or identity-related reasons can in turn lead to the undermining of the common ground on which decisions taken at societal level are based. If this common ground is not clearly identified and shared by all, the public debate becomes distracted from the real issues facing society.

1.2.1.6 Ways forwards

If citizens are to remain genuine actors involved in collective decision-making, the issue of truthfulness in the exchange of information, including on the internet, should become an individual priority. Users must be given the means to protect themselves against polarising mindsets when navigating the internet. In this context, digital literacy can be very useful: simple rules for information literacy can help users navigate and make sense of their social media feeds and other online information sources (Kozyreva et al, 2020).

Digital literacy encompasses the knowledge and skills that the public needs to evaluate the quality of information coming from the media and to distinguish informational content from other types of content – for example, advertisements, opinion content or false information. One aim of researchers is to develop (especially in schools) an information culture, by exploring the media system, the relationships between journalists and other players in society, the role of citizens etc. (Vraga & Tully, 2019; Vraga, Tully & Bode, 2020, Tully et al., 2021). Promoting the skills to correctly interpret numeric information with lower cognitive effort can also help draw attention to key information (Peters, 2017).

There is a wide range of literature on different ways of making people less vulnerable to misinformation through preventive interventions. The effects of interventions based on the use of red flags for unverified or debunked information (Clayton et al., 2020), on crowd wisdom (Pennycook & Rand, 2019; Allen et al., 2021) or on attention-priming techniques (Pennycook et al., 2020), have recently been explored. Researchers' attention has turned in particular to the dissemination of false information on social networks, adapting interventions such as inoculation to these changing environments. Inoculation is a technique introduced by McGuire (1964) and inspired by vaccination, which consists of presenting a poorly argued claim on a subject in order to disqualify any future better-defended argument (Lewandowsky & Van Der Linden, 2021). It even seems to work in modifying already entrenched opinions. Based on this theory, Roozenbeek & van der Linden (2019) developed a game on the topic of fake news. The aim was to make people immune to fake news by putting them in the shoes of a disinformation agent. Fake news seems somewhat less reliable to the participants after they have played the game on social networks. Overall, their results are encouraging but limited.

A form of media education called civic online reasoning (Wineburg & McGrew, 2019; Breakstone et al., 2018; McGrew et al., 2019) has proved to be effective in improving the ability of high school and university students to detect dubious information, particularly on social networks (Wineburg & McGrew, 2016). This approach is based on learning professional fact-checking techniques and is meant to be effective even in cases where disinformation is difficult to detect at first sight. However, such a strategy may fall short in cases where the problem lies in a lower motivation to know the truth on a given issue.

In most situations in our everyday lives, powerful motivational factors shape our relationship to information and direct the way we mobilise our reasoning skills. For reasoning efforts to be aimed at truth, the argumentative context must be collaborative, and individuals must be willing to detach themselves from their own perspective. The context in which communication takes place may in turn make the respective stakes of knowing the truth and maintaining useful beliefs more or less salient. This contextual component is perhaps the most important factor in tackling the challenge of false information on the Internet. A central aspect of citizenship is the possibility of authentic dialogue. In the current political climate, citizens as well as politicians seem to have an increasingly difficult time talking with those holding different opinions about important policy issues (Hess & McAvoy, 2014). For many reasons, a great deal of dialogues we engage in lack the quality of dialogism and are, instead, monological. This means that they end up reinforcing pre-existing views and, oftentimes, strengthening dominant or hegemonic voices. Respect and trust (about others' non-harmful intentions) are also essential features of authentic dialogue – and more generally of democratic settings (Boghossian & Lindsay, 2019) – and they imply listening to others. Reinforcing the ability of individuals to exchange information with others in a respectful way is probably one of the most interesting ways of making them more impervious to manipulation, and less tempted by the superficial use of fallacious arguments.

1.2.1.7 Conclusion

This overview of the literature highlights where and how regulation needs to intervene in the way information is exchanged on the internet, and the role of new tools for recommendations. As far as fake news is concerned, while its potential to cause harm needs to be put into perspective, the dynamics of social networks contribute to blurring the public debate and to undermining the trust placed in institutions and experts, by encouraging the use of information as an argumentative means rather than a way of building a shared and accurate representation of reality. The challenge of regulation and education must therefore focus on the ability of discussion forums to promote respectful dialogue, rather than on combating the limited effects of fake news as such. Further experimental works in this Module will look into how trained bots could foster openness, curiosity and respect, and in turn promote critical thinking.

1.2.2 The problem of individuals' autonomy in their personal choices

1.2.2.1 What is autonomy?

Autonomy of the will, the ability of individuals to decide for themselves the rules of behaviour they adopt, is an ideal that can never be attained. It is intrinsically unachievable because, as human beings, we are subject to physiological constraints beyond our control. Freedom understood as the autonomy of the will[3] is therefore not an original achievement that we can demand to be preserved, but a desirable horizon.

In the late 1970's, the Belmont report[4] had two convictions: the idea that individuals should be considered as autonomous agents (introducing the notion of informed consent); and also, the idea that vulnerable people should be given assistance and protection. However, these two convictions are contradictory, since the individual is considered autonomous on the one hand, and as a patient when he or she is in a state of vulnerability (due to age or a particular condition), on the other. The notion of autonomy is therefore not as self-evident as it first appears, at least in its practical implications.

Discussion of the ethics of AI within experts' committees[5] is set in a context where machines will most probably become increasingly efficient. The aim is thus both to ensure that human beings do not risk abdicating their autonomy in favour of machines, and to empower individuals through technology. The idea behind this is to remove potential constraints on individual freedom, by making people as empowered as possible. That said, it is interesting to think about alternative concepts of autonomy in the context of artificial intelligence. For example, autonomy can also be conceived as the voluntary submission of individuals to the rules that they decide to prescribe for themselves, in accordance with their personal values, preferences and commitment. In this context, it becomes more a question of helping internet users to remain in a position to freely decide what they see, what they accept, what they want, etc.

Beyond the question of whether individuals believe the information to which they are exposed via recommendation algorithms, and whether the latter influence their opinions (we have seen that this is not

[3] This is different from a machine's autonomy, which generally refers to its automatic nature.

[4] The Belmont Report is a foundational document in the field of research ethics, particularly in the context of human subjects research. It was published by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research in the United States in 1979.

[5] E.g., the High Level Expert Group on AI (HLEG AI) set up by the European Commission, the Institute of Electrical and Electronics Engineers (IEEE), and AI4People Institute.

necessarily the case, but that the individual's ability to act as a citizen for the collective good is nevertheless threatened), the use of recommendations algorithms poses other issues linked to autonomy.

1.2.2.2 Does content curation challenge individuals' autonomy?

AI can enable behavioural interventions to be more personalised and contextual for individual consumers by accounting for their unique traits and the specific environments they are in. Combining usage data (created when people use online services) with personal data (relating to an identifiable person), makes it possible to draw up extremely precise files on individuals, their tastes, their interests, and so on. This can lead to interventions that are more effective and sometimes more equitable (Mills et al., 2023). In any case, algorithms are necessary to manage the flow of information to which we are exposed. Every time someone visits a social media news feed, they can be exposed to hundreds if not thousands of stories from friends, pages that they follow, or sponsored content. This equally applies to requests to search engines. There is simply no way we can find anything useful without some form of curation[6]. But the question is, how fine-tuned to people's previous behaviour should the selection be, while still preserving people's autonomy?

One interesting aspect of the question is whether what people’s targeted content influences people's preferences and attitudes. In the domain of consumer product research, some studies have shown that personality-targeted ads can result in more clicks. Others suggest that while online activities (e.g. likes on Facebook) can contribute to developing a psychological profile of users and give insights into people's personalities (Resnick, 2018)[7] personality data minimally makes ads more engaging. In the political domain, while microtargeted ads might reinforce pre-existing viewpoints, evidence suggests they rarely change people's minds or political preferences. The case of Cambridge Analytica's association with the Trump campaign showed mixed outcomes, with some reports suggesting their service didn't provide much value and caused technical issues. All in all, the effectiveness of creating ads tailored to specific personalities remains at least questionable. Even with knowledge of a person's personality, crafting the right ad to be aimed at that person isn't guaranteed to be effective. Advertising is more an art than an exact science, and people are often wrong in their assumptions about what others will find convincing.

Yet, even if micro-targeted content is not particularly efficient when it comes to influencing people's consumer or political choices, some of the challenges of the digital environment include the erosion of autonomy due to manipulative choice architectures, distraction and cognitive overload (Kozyreva et al, 2020). With such precise information on our behaviour, preferences, and attitudes, recommendation systems can recommend the "right" product, the "right" article, the "right" video, at the "right" time, and use the "right" way of doing so, i.e. the one most likely to make us accept them. Algorithms are very effective at selecting content that consumers will perceive as pleasurable and addictive. Analysis of browsing data enables content to be proposed that is close to what users like, but sufficiently different for it to be new. The monetisation of user attention duration on digital platforms incentivizes the deliberate curation of content aimed at capturing and prolonging user engagement. As we saw in the previous section, this strategy often involves the presentation of emotionally evocative, sensational, or highly captivating content to retain user

[6] The diffusion of the printing press generated a comparable panic, with intellectuals worried the world would “fall into a state as barbarous as that of the centuries that followed the fall of the Roman Empire." (Adrian Ballet, 1685).

[7] Based, for instance, on the Big Five personality traits: Agreeableness, neuroticism, openness to experiences, extroversion, and conscientiousness.

Funded by the European Union
KT4D has received funding from the EU’s Horizon Europe research and innovation programme under Grant Agreement no. 101094302.

interest, as certain content resonates more deeply with the human experience due to its relevance to fundamental aspects of life.

So, are social networks addictive? From a scientific point of view, addiction typically includes a compulsive desire, difficulties in controlling its use and, sometimes, a physical withdrawal syndrome. In the case of social media use, there is no such physiological habituation. Social media use falls outside the scope of the DSM-V's recognition of behavioural addictions, limited to video games and gambling. If not classified as a distinct pathology, social networking exhibits nonetheless addictive characteristics by exploiting vulnerabilities in the brain's reward system, a fundamental component of motivation (Sherman et al., 2016). Notably, posting photos that generate numerous likes will trigger the release of dopamine. The strength of this reward, akin to compulsive behaviours observed in Skinner's animal experiments (Krebs, 1983) and analogous to slot machines, intensifies with greater variability and randomness in the reward pattern. Strategies used to foster addiction encompass diverse techniques such as infinite scrolling, notifications, likes, targeted recommendations, etc. These strategies build on priority alerts for cognitive processing and encourage engagement at the expense of ongoing activities.

Manipulative techniques like "dark patterns" are also sometimes used, which prevent users from exercising their autonomy. Some media have for example adopted a "click-bait" model[8], offering headlines that optimise their appeal, regardless of the actual relevance of the content to the user (with headlines such as “9 Out Of 10 Americans Are Completely Wrong About This Mind-Blowing Fact”). Such titles exploit a basic emotional language and arouse curiosity but do not provide enough information without clicking on them. In fact, the stories do not need to be particularly attractive themselves, as long as users are tempted to click on the link. Dark Patterns are interfaces deliberately designed to deceive or manipulate users (Ehrel, 2023; Fitton & Read, 2019). These interfaces are tested on thousands of subjects to determine which adjustments and parameters are most efficient. As Adam Alter says in his book *Irresistible: The rise of addictive technology and the business of keeping us hooked*, “In 2004 Facebook was fun, in 2016 it's addictive" (Alter, 2017). The combination of infinite scrolling and personalised recommendations creates an immersive experience that is similar to a tunnel, inducing a state of absorption and heightened concentration. The intensified perception of control further contributes to the difficulty of disengagement, and exiting the application demands a substantial effort and considerable motivation. These effects are all the more pervasive as the user lacks visual cues: like in the experiment by Wansink et al. (2005) where the use of self-refilling soup bowls altered the feeling of satiety, users of platforms such as TikTok or Instagram no longer feel the need to get out of the tunnel.

Are social networks threatening to psychological health?

While the mental health of young people is deteriorating worldwide, with increasing cases of depression, anxiety and mental disorders (Twenge et al., 2020), the PISA report notes that in almost all the countries studied, loneliness at school has increased. For many experts, social networks are to blame. Between 2009 and 2012, the main social media platforms have undergone radical change, taking on a more toxic dimension: Facebook introduced the 'Like' button, Twitter the 'Retweet' function, and news feeds became algorithmic,

[8] In 2013, the website Upworthy became an internet sensation by utilising this peculiar style for titles which became known as the “Upworthy model.”

based on engagement. As a result, social networks have disrupted social interactions. In France, one child in four feels they spend too much time on their phone, according to a recent survey by the Heaven agency[9]. Yet correlation is not causation, and many other potential factors are at play. Studies on the effects of time spent on social networks are many and contradictory. Some studies conclude that there is a clearly identified effect on mental health, others that there is a weak association, and others that there is no effect at all. A meta-analysis suggests that there is indeed a negative effect, but that it remains very weak (Masciantonio et al., 2023).

However, screen time may not be a good variable. Many factors are confounded in the same analysis, including factors with positive effects on well-being. For example, a Pew Research Center study conducted in 2022[10] shows that half of American teenagers born between 1995 and 2012 feel more accepted online than offline. Does active involvement in online exchanges determine the positive effects of social media? Some studies show that the more involved users are in content production, and the more they interact with other users, the less negative the effect on their mental health. In 2016, a study conducted in Quebec by Morin-Major et al. (2016) showed that the more friends adolescents had on Facebook, the higher their basal levels of cortisol, a hormone produced during stressful experiences. Conversely, the more photos and stories they exchanged, the lower the cortisol level, because the interactions were of better quality. However, other studies (e.g. Beyens et al., 2020) do not find the same results.

The difficulty researchers have in finding consistent results stems from the fact that, more often than not, social networks are not studied per se, but screen time in general, and the population in general. For example, a study conducted by Orben & Przybylski (2019) on hundreds of thousands of British and American teenagers' data revealed a low association between digital technology use and adolescent well-being. But when these same data were analysed by considering only data corresponding to the use of social networks by young adolescent girls, this association was multiplied by four. Adolescent girls (especially between the ages of 11 and 13) are in fact on the front line: one solid finding is that exposure to idealised bodies has a negative impact on body image (Orben et al., 2022). Young girls who undergo social comparison on these platforms are at greater risk of developing psychological disorders (Kleemans et al., 2018).

In conclusion, there are indeed links between social networks and mental health problems, but these depend very much on the individual and the context in which individuals are exposed, and on the characteristics of the platforms (for example, the positivity bias is stronger on Instagram, there are more messages associated with anger on X, there is only video content on TikTok, etc.). A better understanding of the effects of these characteristics, taking into account combinations that turn out to be toxic, will enable better regulation, better adapted to the problems of the users themselves.

[9] https://heaven.paris/files/BORNSOCIAL2023.pdf

[10] https://www.pewresearch.org/internet/2022/11/16/connection-creativity-and-drama-teen-life-on-social-media-in-2022/

Funded by the European Union
KT4D has received funding from the EU's Horizon Europe research and innovation programme under Grant Agreement no. 101094302.

1.2.2.3 How to empower internet users?

The problem with the massive use of recommendation engines in social networks is linked to the question of alignment between the tool and the user. The goals of social networks and platforms are not aligned with those of their users: the former want to make money and keep the latter's attention as much as possible; internet users, on the other hand, simply want to buy items or find information (Russell, 2019) - and more generally speaking, maximise their well-being. As people are naturally vigilant about attempts to manipulate them (Sperber et al., 2010), the question of the transparency of the intentions of those who offer online content becomes a critical issue. If people know what to expect, they will naturally be more vigilant. Consequently, it becomes imperative to comprehend the mechanisms underlying digital networks and cultivate a foundational knowledge of their operations, emphasising the significance of media literacy. The opacity surrounding the authentic motives of private entities and the inscrutability of algorithms contribute to an illusion of freedom and choice for individuals. While ostensibly free to navigate the Internet according to personal preferences, online choices made by individuals often contradict the notion of a conscious and autonomous self.

Most solutions proposed by the same platforms remain inadequate in addressing the issue at hand. For instance, widespread acceptance of cookies by the majority may be more a result of limited alternatives than a conscious choice. Rather than empowering individuals, terms and conditions often demand the assent of users without providing any space for deliberation. Intricate language and extended document lengths can also contribute to creating an environment that lacks transparency and leaves users ultimately uninformed.

Getting back in control

How can we reclaim a sense of algorithmic sovereignty? Traditional transparency measures, such as tools provided by platforms like Facebook's "Why am I seeing this?" features are inadequate as they offer only superficial information and rely on active user engagement to be effective.

Making internet users more autonomous can be partly achieved by increasing information literacy and cognitive resistance to manipulation. A valuable strategy for promoting autonomy in the context of social media usage involves the implementation of feedback mechanisms, such as tracking app usage time, setting alerts, and managing notifications. Such control tools can effectively mitigate excessive engagement, and they can also be incorporated into educational curricula, ensuring that individuals, particularly the youth, are equipped with the knowledge to navigate social networks responsibly. Psychological science can further contribute to addressing these challenges by informing the design of cognitive tools and interventions. Different types of interventions have been explored to foster better online decision-making and resist manipulation. Some argue for a psychological *boosting* approach, aimed at enhancing agency in digital environments and at fostering users' reasoning and resilience to manipulation. In this approach, individuals are inoculated against manipulation by enhancing their competence to detect such ads and make informed decisions. For instance, Lorenz-Spreen et al. (2021) showed that a short, simple intervention—prompting participants to reflect on their own personality—can significantly increase people's ability to accurately identify ads that were targeted at them. Personal reflection and feedback significantly improved the ability to correctly identify targeted ads.

In any case, more regulation is needed, that would promote transparency in the intentions of online platforms and services - and not just in terms of getting users to sign agreements. Regulations should take into account the psychological constraints of individuals by, e.g., questioning the genuine agreement of users with the terms and conditions in GDPR online forms. This approach would align regulations with users' cognitive and emotional considerations, enhancing the transparency and ethical implications of algorithmic choices in digital environments.

The experimental work carried out in this project will focus precisely on users' intuitions about their sense of autonomy in various situations involving content curation.

1.2.2.4 Conclusion

This second section looked at the problem posed by architectures designed to manipulate individuals on the Internet, by concealing, as far as possible, the intentions of online service providers. On the one hand, a certain form of content curation is necessary, and sometimes desired by individuals. On the other hand, by placing them in situations where their autonomy is confiscated, and even though they have voluntarily put themselves in this situation, users end up making choices that go against their own interests. Efforts in terms of both education and regulation must ensure a better match between the interests of individuals and the objectives they set for themselves on the one hand, and their interactions in a world governed by content derived from AI-based algorithms on the other.

1.3 Methodology to advance the state of the art

The experimental component of Module A explores individuals' intuitions on the compatibility between algorithms and what should be individuals' autonomy. It will further characterise internet users' behaviours when faced with online choices potentially undermining their autonomy, and test potential ways forward.

Cross-cultural survey on individuals' intuitions

We are designing a survey modelled after the Moral Machine Survey (Awad et al., 2018)[11]. The survey questions will be structured around manipulated scenarios, drawing insights from the KT4D participatory design sessions, and incorporating diverse cultural backgrounds in our data collection. Within this module, the items will serve to answer the following questions: 1. What are people's intuitions regarding autonomy when an internet user is offered content that they can hardly refuse? 2. What would be the moral value of empowerment techniques: (under what conditions) would people consider "boosts" vs. "nudges”[12] appropriate solutions?

Behavioural experiments on critical thinking and autonomy

We will also rely on techniques from the field of experimental psychology to better characterise how people evaluate AI-generated information and/or content selected through AI-based algorithms. Are people more vigilant when they know how AI works and/or when they are aware of the intentions of those who provide them with this content? We will study the evaluation of information with manipulated sources (AI vs. Expert vs. Layperson), as well as manipulated levels of accuracy and relevance. The stakes associated with the truthfulness of information will also be measured and included as a variable for assessing the propensity to exercise critical thinking.

[11] http://moralmachine.mit.edu/

[12] While nudges coopt people’s existing cognitive biases to affect behavioural changes, boosts train people in employing existing decision heuristics or employing new ones (Grüne-Yanoff, 2018).

Funded by the European Union

KT4D has received funding from the EU’s Horizon Europe research and innovation programme under Grant Agreement no. 101094302.

Regarding online choices, we will clarify what consent means in situations when people are asked to agree on terms and conditions. We will monitor the micro-decisions people make in such situations and examine the contextual factors that matter most. Moreover, we will look into whether receiving tailor-made recommendations from algorithms make individuals less open to diverse contents.

Training online users via bot interactions

A final objective will be to test the efficiency of a particular approach for promoting individual autonomy and critical thinking. We will investigate the efficacy of exposing individuals to pre-trained bots, with a focus on fostering open-mindedness and reducing susceptibility to toxicity.

Through this comprehensive approach, this module will not only inform on individual perceptions but also contribute to the development of concrete strategies that prioritise sound ethical considerations: we will derive the conditions for an ethical system capable of addressing their concerns.

1.4 Interim conclusion: the conditions of a truly empowering system

This literature review identifies key areas where regulation is essential in shaping internet information exchange and the use of new recommendation tools. It emphasises the role of social network dynamics in altering the public debate and eroding trust in institutions. Rather than combating fake news per se, it suggests that the focus should be on regulating and educating discussion forums to encourage respectful dialogue and foster shared and accurate representations of reality. Additionally, it addresses the challenge of architectures designed to manipulate (sometimes in good faith) individuals on the internet, emphasising the need for a balance between necessary content curation and the preservation of user autonomy.

This module will clarify, on the basis of the point of view of the users themselves, what information or what skills people need to avoid falling into the rabbit hole of manipulative techniques designed to optimise user engagement. And by contrast, which regulations are necessary when the burden on users becomes too heavy.

1.5 References

No	Description/Link
R1	Acerbi, A. (2019). Cultural evolution in the digital age. Oxford University Press.
R2	Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. Science advances, 7(36), eabf4393.
R3	Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. Social Media+ Society, 9(1), 20563051221150412.
R4	Alter, A. (2017). Irresistible: The rise of addictive technology and the business of keeping us hooked. Penguin.
R5	Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. Nature, 563(7729), 59-64.
R6	Benkler, Y., Faris, R., & Roberts, H. (2018). Network propaganda: Manipulation, disinformation, and radicalization in American politics. Oxford University Press.
R7	Beyens, I., Pouwels, J. L., van Driel, I. I., Keijsers, L., & Valkenburg, P. M. (2020). The effect of social media on well-being differs from adolescent to adolescent. Scientific Reports, 10(1), 10763.
R8	Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, 114(28), 7313-7318.
R9	Broockman, D. E., & Green, D. P. (2014). Do online advertisements increase political candidates' name recognition or favorability? Evidence from randomized field experiments. Political Behavior, 36, 263-289.
R10	DeVito, M. A. (2017). From editors to algorithms: A values-based approach to understanding story selection in the Facebook news feed. Digital journalism, 5(6), 753-773.
R11	Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes?. The International Journal of Press/Politics, 26(1), 69-91.
R12	Dunbar, R. I. (2012). Social cognition on the Internet: testing constraints on social network size. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1599), 2192-2201.
R13	Elswah, M., & Howard, P. N. (2020). "Anything that causes chaos": The organizational behavior of Russia Today (RT). Journal of Communication, 70(5), 623-645.
R14	Erhel, S., Drouard, J., Jacob, F., Lumeau, M., Suire, R., & Gonthier, C. (2023). Predictors of problematic internet use in the everyday internet activities of a French representative sample: The importance of psychological traits. Computers in Human Behavior, 108099.
R15	Etienne, H. (2021). The future of online trust (and why Deepfake is advancing it). AI and Ethics, 1(4), 553-562.
R16	Fitton, D., & Read, J. C. (2019). Creating a framework to support the critical consideration of dark design aspects in free-to-play apps. In Proceedings of the 18th acm international conference on interaction design and children (pp. 407-418).

R17	Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on Twitter?. In Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science (pp. 179-192).
R18	Graeupner, D., & Coman, A. (2017). The dark side of meaning-making: How social exclusion leads to superstitious thinking. Journal of Experimental Social Psychology, 69, 218-222.
R19	Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. Science, 363(6425), 374-378.
R20	Grüne-Yanoff, T. (2018). Boosts vs. nudges from a welfarist perspective. Revue d'économie politique, (2), 209-224.
R21	Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive–motivational mechanisms of political polarization in social-communicative contexts. Nature Reviews Psychology, 1(10), 560-576.
R22	Kalla, J. L., & Broockman, D. E. (2018). The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. American Political Science Review, 112(1), 148-166.
R23	Kleemans, M., Daalmans, S., Carbaat, I., & Anschütz, D. (2018). Picture perfect: The direct effect of manipulated Instagram photos on body image in adolescent girls. Media Psychology, 21(1), 93-110.
R24	Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. Psychological Science in the Public Interest, 21(3), 103-156.
R25	Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National academy of Sciences of the United States of America, 111(24), 8788.
R26	Lantian, A., Wood, M., & Gjoneska, B. (2020). Personality traits, cognitive styles and worldviews associated with beliefs in conspiracy theories. Routledge handbook of conspiracy theories, 155-167.
R27	Lewandowsky, S., Ecker, U. K., Cook, J., Van Der Linden, S., Roozenbeek, J., & Oreskes, N. (2023). Misinformation and the epistemic integrity of democracy. Current Opinion in Psychology, 101711.
R28	Lewandowsky, S., Robertson, R. E., & DiResta, R. (2023). Challenges in understanding human-algorithm entanglement during online information consumption. Perspectives on Psychological Science, 17456916231180809.
R29	Lorenz-Spreen, P., Geers, M., Pachur, T., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Boosting people’s ability to detect microtargeted advertising. Scientific Reports, 11(1), 15541.
R30	Lynch, M. P. (2004). Who cares about the truth. Chronicle High. Educ, 51, B6.
R31	Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. Cognitive research: principles and implications, 5, 1-20.
R32	Masciantonio, A., Résibois, M., Bouchat, P., & Bourguignon, D. (2023). Social Network Sites and Well-Being: Is it Only a Matter of Content?. International Review of Social Psychology, 36(1), 6.

R33	Mercier, H. (2020). Not born yesterday: The science of who we trust and what we believe. Princeton University Press.
R34	Mills, S., Costa, S., & Sunstein, C. R. (2023). AI, Behavioural Science, and Consumer Welfare. Journal of Consumer Policy, 46(3), 387-400.
R35	Morin-Major, J. K., Marin, M. F., Durand, N., Wan, N., Juster, R. P., & Lupien, S. J. (2016). Facebook behaviors associated with diurnal cortisol in adolescents: Is befriending stressful?. Psychoneuroendocrinology, 63, 238-246.
R36	Morisseau, T., Branch, T. Y., & Origgi, G. (2021). Stakes of knowing the truth: a motivational perspective on the popularity of a controversial scientific theory. Frontiers in Psychology, 3800.
R37	Nyhan, B. (2018). Fake News and Bots May Be Worrisome, but Their Political Power Is Overblown. International New York Times.
R38	Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. Nature human behaviour, 3(2), 173-182.
R39	Orben, A., Przybylski, A. K., Blakemore, S. J., & Kievit, R. A. (2022). Windows of developmental sensitivity to social media. Nature Communications, 13(1), 1649.
R40	Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. Proceedings of the National Academy of Sciences, 118(26), e2024292118.
R41	Resnick, B. (2018). Cambridge Analytica's 'Psychographic Microtargeting': What's Bullshit and What's Legit. Vox, March, 26.
R42	Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. Proceedings of the National Academy of Sciences, 118(50), e2102141118.
R43	Schaffner, B. F., & Luks, S. (2018). Misinformation or expressive responding? What an inauguration crowd can tell us about the source of political misinformation in surveys. Public Opinion Quarterly, 82(1), 135-147.
R44	Sherman, L. E., Payton, A. A., Hernandez, L. M., Greenfield, P. M., & Dapretto, M. (2016). The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. Psychological science, 27(7), 1027-1035.
R45	Trivers, R. (2006). Reciprocal altruism: 30 years later. Cooperation in primates and humans: Mechanisms and evolution, 67-83
R46	Twenge, J. M., Haidt, J., Joiner, T. E., & Campbell, W. K. (2020). Underestimating digital media harm. Nature Human Behaviour, 4(4), 346-348.
R47	Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. Trends in cognitive sciences, 22(3), 213-224.
R48	Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.
R49	Wansink, B., Painter, J. E., & North, J. (2005). Bottomless bowls: why visual cues of portion size may influence intake. Obesity research, 13(1), 93-100.

2 Module B: AI, trust and awareness

2.1 Introduction

In this section, we will examine the question of the conditions for trust in systems, people and ultimately institutions relying on AI. The aim of regulation must be to strengthen trust in institutions, a basic condition for the functioning of democracy. Alongside inclusivity and fairness, political sovereignty is also at stake. Proposals of ethics committees should be carefully examined in the light of the true point of view of the individuals, and the customs and habits of the societies in which they live.

From this point of view, regulating the way social networks operate is particularly important. We've seen the risks for the quality of information, and thus to the quality of public debate. There is a duality in the influence of social media platforms on democratic systems: they provide both a very efficient forum for expression, but also undermine the trust placed in institutions that usually guarantee the reliability of information, as the origin of information becomes increasingly blurred.

Another aspect is the relationship between citizens themselves. The advent of artificial intelligence in everyday life has promoted modes of exchange and interaction, to which all sectors of society must adapt. The question we are exploring here concerns the nature of trust and distrust as applied to AI, both in the production of informational or artistic content (via tools such as Chat-GPT or Midjourney), and in the selling of online content (via the exploitation of big data).

We will attempt to answer this question from a particular angle, drawing on the psychological literature on moral intuitions, and examining how these intuitions operate in various situations affected by AI. A series of interconnected questions will be raised. Firstly, there is a need to explore the moral considerations surrounding the integration of tools such as ChatGPT or Midjourney in educational and professional settings, by looking into the conditions that define its acceptability and the impact on perceptions of those relying on AI for idea and artefact generation. The second part of this section focuses on the evaluation of artificially generated ideas or works, which can raise concerns about the preservation of human values and skills. Finally, we will look at the issue of trust in the benevolence of web platforms and see where the need for greater transparency in the collection of personal data is most pressing.

2.2 State of the art: literature review

2.2.1 AI and social fairness

In the course of their evolution, humans have organised themselves in cooperative settings, and it was crucial to share cooperation benefits in a mutually advantageous manner (Baumard, 2011). Those who were too individualistic risked losing partners, while those taking less than their share risked exploitation by partners receiving more than they contributed. This competition resulted in the evolution of a sense of fairness, a cognitive adaptation aiming to equally share cooperation benefits. Evolution favoured fairness because impartial individuals were chosen more often as cooperation partners. Morality may thus have evolved as an adaptation to an environment marked by competitive dynamics among individuals seeking selection and recruitment in mutually advantageous cooperative interactions (Baumard, André, and Sperber, 2013). Being fair and reliable made individuals attractive collaborators, leading to the widespread acceptance of fairness in groups. Thus, people have a natural sense of what is right in distributing benefits, similar to making mutually beneficial agreements. Such behaviour often follows a pattern like social contracts, even in short-term interactions.

The issue of equity, defined as rewarding according to contribution, is a primary concern in the context of AI. In particular, equity is a key component of human fairness, as established in both philosophical discourse and scientific inquiry (Debove et al., 2017). Philosophers have long underscored, since Aristotle, the importance of the notion that greater effort warrants greater benefits (Konow, 2003; Skitka, 2012). Psychological research, particularly on distributive justice and equity theory, provides substantial empirical backing for the importance of merit-based rewards (Homans, 1958). A consistent finding is that receiving more or less than one deserves results in distress, prompting efforts to restore equity by adjusting one's contribution (Adams & Jacobson, 1964). Preferences for income distributions reflecting strong work-salary correlations, greater rewards for more valuable contributions, and overall meritocratic distributions are evident in both micro- and macro-justice contexts (Baumard et al., 2013). This body of literature serves as a valuable foundation for examining the implications of tools like ChatGPT in diverse social contexts, such as education and work.

The adoption of AI-based tools is problematic in that it creates more occasions of potential inequality and dishonesty, eroding trust between group members who expect equal treatment for all. In academic contexts, the deployment of AI-tools such as ChatGPT raises concerns related to bias and fairness, as well as plagiarism. For instance, according to Reich (2022), AI's ability to generate high-quality essays may erode academic integrity. As AI technologies become easily accessible and increasingly difficult to detect, the potential for students to misuse these tools to cheat on assignments increases.

On the other hand, the use of AI tools like ChatGPT should not necessarily be seen as cheating or plagiarism. They may involve a process of analysis and evaluation from the student's side, implying that they are doing a work of their own. What, then, are the conditions that would enable these tools to be used with equanimity, so that they are not associated with a threat linked to a situation experienced as potentially unfair?

To answer this question, research into students' perceptions of what constitutes cheating in the academic context, and the motivations behind it, is quite relevant. For example, in a study by Beasley (2014), students reported that increased awareness of what constitutes cheating and academic dishonesty might have prevented their actions (e.g., through explicit guidelines during exams and assignments; clearer instructions from their professors, etc.). Many students rationalise their cheating by blaming professors, accusing them of failing to emphasise academic integrity or not creating a learning environment that supports honesty. Noteworthily, the use of AI in academic context creates a vacuum that is seen as an opportunity for some, and as an unfair advantage for others.

Broadly changing academic conduct policies to classify AI as a source of plagiarism would be a hasty decision and would be tantamount to throwing the baby out with the bathwater. Both students and teachers should develop 'AI literacy', which includes being aware of AI's pervasiveness, gaining the ability to use it, recognising that everyone can use it, and applying critical thinking to the content produced by AI. Rather than prohibiting AI tools in academia, it could be more beneficial to educate students about AI, its usage, and ethical implications. But whether this means preventing the use of AI-tools in certain test conditions, or making use of them by adapting what is asked of students, in the end, it is always a question of reassuring the institution's ability to sanction students fairly.

2.2.2 AI and the valuation of human productions

Achieving with great effort what a machine can do in just a few minutes may seem pointless. The purpose of such effort may be considered differently when it comes to developing a particular skill. In this case, the value lies in the intellectual effort required to achieve that goal. People can be replaced for a task they don't want to do themselves, but personal effort is key to progress. Yet certain skills may seem outdated, such as manual arithmetic in the age of calculators. While basic manual calculation skills are still relevant, it makes more sense to focus on learning how to use these machines for advanced mathematical applications.

The case of cover letters is one of many examples. Knowing how to write the first draft of a cover letter using a tool such as ChatGPT can be just as useful as writing one from scratch. The ability to write a cover letter independently seems to have lost some of its importance. A large proportion of writing tasks in general are moving towards activities such as editing, proofreading and rewriting. Moreover, the increasing power of linguistic models is likely to accelerate this transformation. As a result, our approach to writing is likely to undergo a radical change.

With the advent of AI, the question of what is important to value as skills and talents is being raised more than ever. Research on how people evaluate the value of productions (made by humans or by AI), and art in particular, provides an interesting insight, highlighting what is actually valued.

For instance, Magni et al. (2023) showed that people evaluate creativity differently based on whether artefacts are thought to be produced by either AI or humans. They found that there is a context-dependent bias against AI in creative production. Specifically, the bias against AI creative artefacts manifested in some contexts (like paintings) but not others (like advertisement posters and business ideas). A key finding was that AI is perceived to exert less effort than humans in the production of creative outputs, and this perception mediates the relationship between the producer's identity and the creativity attributed to the production. Thus, effort perception is a significant mediator in the evaluation of creative products. A folk psychology framework allows us to understand how the identity of the producer impacts people’s evaluations of creative works: humans might perceive artificial agents to lack emotional and intentional capabilities, which are seen as important in creative processes.

Another study (although not on the subject of artificial intelligence) highlights the role of effort in assessing artistic production. Kruger et al. (2004) tested the hypothesis that people used the effort invested in the creation of a product as a heuristic, or mental shortcut, for judging its quality. The idea was that this heuristic is employed because direct assessment of quality can often be difficult. By manipulating the reported effort put into the creation of various artefacts (like poems, paintings, and suits of armour) and keeping the actual quality constant, the researchers sought to demonstrate that judgments of quality could be swayed erroneously by the perceived effort alone, potentially leading to misjudgment. They made a distinction between self-generated effort, which is associated with dissonance theory, and other-generated effort. The effort heuristic discussed in this paper refers to the judgments about efforts made by others, not by oneself. The authors suggest that the reliance on the effort heuristic is stronger in situations where the quality of the object being evaluated is ambiguous. They argue that in the case of objects like art, where quality is not readily apparent and can be subjective, the reported effort put into the object's creation becomes a default indicator of quality. The effort heuristic is widely used among different types of individuals (including self-identified experts) and across domains, indicative of its general appeal as a basis for judgement. So while effort can indeed be a generally reliable indicator of quality, there are instances where it may fail to accurately reflect the true value of a work, leading to incorrect judgments of quality.

Other studies show that humans tend to have a preference for artworks labelled as human-created over those labelled as AI-created (Bellaiche et al., 2023). This observed preference is consistent across several judgement criteria, such as Liking, Beauty, Profundity, and Worth. Labels such as "Human-created" or "AI-created" can heavily influence individuals' enjoyment, trust, and valuation of the artworks, regardless of the actual creator. Interestingly, higher perceived effort leads to a greater appreciation when labelled as "human-created." Similar findings were observed in the music domain. There seems to exist an "AI composer bias" where listeners' enjoyment and quality ratings of music are influenced by their knowledge or beliefs about the composer's identity (Shank et al., 2023). People like music less and judge it as lower in quality if they believe it was composed by an AI compared to believing it was composed by a human. Even with no actual difference in the music itself, the perceived identity of the composer can significantly affect listeners' aesthetic judgments.

Intentionality (i.e. the intention of the author of an artwork) seems to be a fundamental ingredient in assessing the quality of an artistic production.

As art is not only a physical stimulus, but also a means of human expression and communication, the importance attached to the author's true intentionality can outweigh aesthetic assessment. Creation indeed involves conceiving a mental concept for the artefact (by an author) and then physically manufacturing it (by an instrument), typically within a single individual. The concept of Mental Primacy (Judge et al., 2020) emphasises that the generation of ideas (mental labour) is valued more than the execution of those ideas through physical labour, in Western societies. Intellectual Property laws reflect this by protecting an author's mental concepts, sometimes even before their physical realisation. There is a belief that mental and physical processes complement each other in the creation process, producing a more valuable artefact. This notion posits that a material artefact's true value is released when an idea is embodied through skillful and effortful physical work (Mind-Body Complementarity). Both the mental and physical essences of makers are transmitted to the material artefact. This influences how artefacts are valued and perceived in terms of authenticity, ownership, and connection to the creator. Folk theories of artefact creation significantly impact human-artefact relations, affecting how people evaluate different types of artefacts and their makers. Such theories can shape cultural materials and influence societal views on production and consumption. Other cultures have different preconceptions about value and authenticity (see for instance Coleman, 2001, who shows that paintings considered inauthentic from a Western perspective may be authentic within Aboriginal contexts).

The perceived value of an artwork is heavily influenced by whether an object is categorised as art, which can be shaped by the creator’s intent, and not just by its functional use (Newman and Bloom, 2012). People's judgments about the value of art are impacted by the intentions of the original manufacturer to make the object a piece of art or an artefact. This ties in with the notion of artistic performance being inherently tied to the artist's intent and the unique creation process.

In the economic sphere, consumers have a special appreciation for handmade products, viewing them as more attractive compared to machine-made products. This perceived value is known as the "handmade effect.". One reason for the handmade effect may be that consumers believe handmade products are imbued with love (Fuchs et al., 2015). This belief would come from the perception that artisans invest personal care and affection into the production process. Despite rapid technological advancement in manufacturing, human craftsmanship is unlikely to become obsolete. Handmade products fulfil a consumer desire for goods that embody human attributes like love and care, which cannot be replicated by machines. Skills associated with hand making products are thus likely to remain valuable in the labour market due to consumer preferences for goods made with human touch. This is an interesting parallel to be drawn with the case of AI-generated content: one can hardly imagine that human creation will lose its value anytime soon. The question is rather to understand how people attribute value to human creation.

2.2.3 AI and the benevolence of institutions

The deployment of AI is likely to give rise to situations of injustice, particularly through the use of opaque algorithms, which calls for greater transparency in these systems. It is not only about addresssing social disparities and ensuring fair access to AI-related services, but also about preventing individuals from being unfairly treated by AI. This section aims to explore some key considerations that influence individuals' trust in the benevolence of web services that collect personal data.

For many experts, the development and use of large language models in natural language processing is leading to significant risks for society, including the propagation of biases, perpetuation of harmful stereotypes, and the production of misleadingly fluent but potentially harmful content without accountability (Bender et al., 2021). While some algorithms are indeed biased, due to the data on which they have been trained, the overall prevalence of bias remains relatively limited (relative to the base frequency). Greater risks lie in the use of discriminating categories, and in unfounded reliance on AI tools, rather than in biases inherent in the algorithms themselves. The temptation to call for absolute transparency is strong, but one might in fact wonder to what extent is such transparency necessary, or even desirable. The question of how and to whom this transparency could be really useful is not trivial. There is in fact a certain incompatibility between clarity and understandability, because good explanations often require the use of intuitive but misleading representations. Rather than making absolute demands in terms of transparency, we need to consider the conditions for informed confidence on the part of users in AI-based services.

2.2.3.1 Personal data collection

The concept of personal data is somewhat confusing. Intimacy, privacy, personal data: all these words are not equivalent. What should really be protected? Recommendations can hardly be limited to a simple and unique claim for universal protection, which would apply to all private information, whatever its nature and the context in which it is used. Instead, we need to discern what exactly is meant by personal data. Violations of intimacy must obviously be condemned. But what about private life, understood as the totality of ordinary daily activities?

The history of the notion of privacy traces a complex evolution from ancient Greece to the 21st century. In ancient Greece, the distinction between private and public spaces shaped society, with the collapse of democratic institutions marking the loss of the complementary functions of these spaces. The Middle Ages saw the emergence of networks of solidarity and shared resources, challenging the concept of private and public realms. The modern age, particularly in England and France during the Revolution, introduced liberties that protected citizens from public interference in personal affairs. The 20th century witnessed totalitarian regimes aspiring to control individuals' private spheres, leading to the inclusion of privacy protection in the Universal Declaration of Human Rights in 1948. Authoritarian states exploited personal data to maintain their hold over those they dominate, and totalitarian states exploited it even more.

For Hannah Arendt, the opening up of the private sphere to the public gaze - and therefore, in a way, its disappearance as private - leads to a shrinking of the common sphere, and therefore of the political sphere (Biesta, 2012). The more private communities and solidarity networks provide for individuals, the less important the public sphere becomes, because decisions are taken elsewhere, and the space for politics itself is accordingly reduced. And the more the private sphere invades the public sphere, the more paralysed the discussion of public affairs inevitably becomes.

The 21st century now presents new challenges. Web platforms are mining vast amounts of information about the world, collecting individual data for profit, challenging the traditional roles of states in security, education and health. This development reflects an ongoing struggle between the protection of privacy and the risk of intrusion by states and by corporations. As the case of China illustrates, the risk of totalitarian states remains, underlining the relevance of privacy protections in the face of changing societal and technological landscapes.

Furthermore, the disclosure of personal information grants real power over individuals. Maintaining a representative democracy requires that privacy be protected, so that individuals can freely express their opinions and contribute to public debate, without having to bow to central control. Personal security also requires that individuals are not vulnerable to malicious private actors and hackers.

That said, the exploitation of personal data by governments does not necessarily lead to an abuse of power. Governments also use large quantities of personal data for the public good, in the exercise of their duty to protect citizens. Provided that they do not take advantage of it to use it for coercion. Democratic safeguards are therefore needed to guarantee the trustworthiness of our institutions. Citizens must weigh up the risks of sharing their personal data against the benefits, while guaranteeing the transparency and reliability of our institutions.

International regulations such as GDPR provides a framework for data exploitation in Europe, including by limiting the collection of information to what is strictly necessary, securing access to databases, destroying data after a certain period of time, encouraging anonymisation, and so on. But the relevance of these regulations depends on the context. The protection of personal data, presented as a universal requirement, does not refer to the same thing depending on whether the data is private, intimate, anonymous or not, whether it is held by a government or by a private company, and so on. A state can use health information about its population to protect it from a pandemic, as was the case during the Covid crisis. It can also use Big Data and AI for much more dubious purposes, such as determining an individual's political orientation (Rasmussen, Ludeke & Klemmensen, 2023; Kosinski, 2021).

Can individuals trust that companies prioritise their personal interests, even when profit is their primary goal? This inquiry relates to the dynamics of trust, probing the balance between corporate profit motives and genuine consideration for the well-being of users in an interconnected and data-driven world. As Lewandowsky and Pomerantsev (2022) observe, there is a power imbalance between platforms who have vast amounts of data on users (and can thus manipulate content delivery) and users who often have little understanding of how their data are being used or how algorithms dictate their content feed.

2.2.3.2 Can we trust the benevolence of private platforms relying on AI-based algorithms?

In recent years, online platforms have been showing increased recognition of the issues surrounding the algorithms used to maximise user engagement and have committed to working on solutions to the problem of the problematic use of social media among the youth. For instance, TikTok introduced in 2023 a 60-minute time restriction for users under 18, which requires the user to enter a password to continue. Disabling push notifications has been made easier, and notifications are now deactivated after 9pm for users aged 13-15, and after 10pm for those aged 16-17. In 2022, Instagram had introduced a feature called "take a break", as another measure to address user engagement and well-being.

In France, a law was passed in July 2023 to introduce the digital age of majority: under the age of 15, parental consent is required for young people to access social networks. The same year, the European Digital Service Act was passed to regulate the activities of platforms and provide greater protection for users. Platforms now have to assess the risks induced by their recommendation systems and mitigate them. They have to explain in the Terms and Conditions of use, the parameters used and the options that are available for changing them. They also have to offer the possibility of an alternative news feed (although this may remain an option that the user must activate). The DSA also requires platforms to provide transparency on advertisements so that users know that what they are receiving is an advertisement, on whose behalf it is sent, and why the advertisement was selected for them.

Then again, it has to be for the benefit of users, and not only to protect against legal risks. In other words, it must make a real difference to users. Regulation has to be useful and desirable, and the user has to know why a given regulation is better for the majority of users. Ideally, platforms should allow third parties to provide recommendation systems that would allow users to select algorithms that suit them, providing them with interesting information while at the same time, making that choice genuinely informed.

Opaque algorithms and secret data

User data is also very closely guarded. Even public data (for example, the number of tweets for a keyword) is very difficult to manipulate, and much of it is only accessible to authorised parties. It is therefore very difficult to assess the effects of these algorithms. TikTok has launched its own API[13] to analyse the activity of the platform's users, but it really provides very little data. Furthermore, there is no guarantee that what the platforms share is actually up to date. And in any case, most of the algorithm's rules are derived from AI models trained on large amounts of behavioural data. The system is emergent, and only marginally dependent on rules as explicit as “If the post is toxic, then make it more visible.”

Communities of researchers - such as AI Forensics' TikTok Global Observatory[14] - get around the issue by running simulations. Bouchaud, Chavalarias & Panahi (2023) showed that Twitter's "For You" recommendation feed contained 50% more toxic messages than the feed of accounts that the user followed. But each new version of the applications makes it more difficult for researchers to capture what lies underneath social media platforms’ activity. The DSA could facilitate this approach by allowing researchers to get access to protected data (Article 40). However, such access will need to be highly controlled, and raise additional questions related, e.g., to the process of authorising access and assessing the quality of the

[13] API stands for Application Programming Interface, which is a set of functions, procedures, methods or classes used by computer programs to request services from the operating system, software libraries or any other service providers running on the computer (source: Wikipedia.com)

[14] https://tkgo.aiforensics.org/

associated research. Another solution for researchers is to rely on data donation, by asking participants recruited for this purpose to install extensions on their computers to collect their data. This is a time-consuming and cumbersome process, but which remains a reliable way of obtaining reliable information.

2.3 Methodology to advance the state of the art

The experimental component of Module B explores affective and social dimensions of trust applied to systems, people and institutions relying on AI.

Cross-cultural survey on individuals' intuitions

We are designing a survey modelled after the Moral Machine survey[15]**.** The survey questions will be structured around manipulated scenarios, drawing insights from the KT4D participatory design sessions, and incorporating diverse cultural backgrounds in our data collection. Within this module, the items will serve to answer the following questions:

1. Under what conditions (if at all) is the integration of tools such as ChatGPT at school or at work considered morally right? How do we evaluate those who rely on AI to produce something? What are the key issues that matter to people?

- Is it about fairness/inequality? Does the problem lie in the difficulty to assess others' skill, and the true merits of each member of society?
- What is the role of transparency of intentions with respect to how a good or a service was created? Is it above all a matter of knowing the intentions behind the use of AI? (i.e., people don't want to be fooled).
- What information should people be provided with when using a good or service potentially produced with AI? What levels of trust are actually placed in companies, researchers, politicians, etc. if we suspect them to rely on AI to do their job?

2. How do people assess the value of an artificially generated idea or work?

3. What are the human values and skills that people feel to be most important? Is it about intentionality and the experience of being a human? About creativity or effort? How would a "made by human" label help foster trust in the quality of both AI and human production?

4. In what circumstances is it acceptable for individuals that private parties have access to their personal data?

Behavioural experiments on critical thinking and autonomy

We will also rely on experimental designs from the field of behavioural economy. Situations in which people will be in the position to reward other agents, depending on their use of AI to produce either informational or artistic content. We will manipulate the agents' choices, whether it was open or hidden, as well as the social context of the interaction (e.g., secure vs. insecure).

[15] http://moralmachine.mit.edu/

2.4 Interim conclusion: the conditions of a trusting society

The literature surveyed explored domains wherein the utilisation of artificial intelligence (AI) poses challenges to the trust dynamics among diverse components of society. An overarching query emerges: What prerequisites must be met to ensure that the application of AI contributes to the preservation of institutional trust? Examining AI within the context of social fairness, particularly in the realm of academic integrity and the challenge of plagiarism, underscores the imperative of a transparent framework for individual assessment when employing such tools, as their integration sparks myriad concerns. We then considered the interrogation prompted by AI-generated outputs, prompting contemplation on the fundamental values inherent in human production. Finally, the issue of trust in the benevolence of institutions was examined.

This module serves as a substantive contribution to the Social Risk Toolkit, furnishing pivotal insights into the strategic areas necessitating educational and regulatory interventions to effectively address the challenges confronting citizens.

1.

2.5 References

No	Description/Link
R1	Adams, J. S. (1963). Towards an understanding of inequity. The journal of abnormal and social psychology, 67(5), 422.
R2	Adams, J. S., & Jacobsen, P. R. (1964). Effects of wage inequities on work quality. The Journal of Abnormal and Social Psychology, 69(1), 19.
R3	Anders, B. A. (2023). Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking? *Patterns, 4(3).*
R4	Baumard, N. (2011). Punishment is not a group adaptation: Humans punish to restore fairness rather than to support group cooperation. *Mind & Society, 10*, 1-26.
R5	Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36(1),* 59-78.
R6	Beasley, E. M. (2014). Students reported for cheating explain what they think would have stopped them. *Ethics & Behavior, 24(3),* 229-252.
R7	Bechwati, N. N., & Xia, L. (2003). Do computers sweat? The impact of perceived effort of online decision aids on consumers' satisfaction with the decision process. *Journal of Consumer Psychology, 13(1-2),* 139-148.
R8	Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., ... & Seli, P. (2023). Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research: Principles and Implications, 8(1)*, 42.
R9	Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
R10	Biesta, G. (2012). Becoming public: Public pedagogy, citizenship and the public sphere. Social & Cultural Geography, 13(7), 683-697.
R11	Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition,* 181, 21-34.
R12	Bouchaud, P., Chavalarias, D., & Panahi, M. (2023). Crowdsourced audit of Twitter's recommender systems. Scientific Reports, 13(1), 16815.
R13	Coleman, E. B. (2001). Aboriginal painting: identity and authenticity. *The Journal of Aesthetics and Art Criticism, 59(4),* 385-402.
R14	Debove, S., Baumard, N., & André, J. B. (2017). On the evolutionary origins of equity. *PLoS One, 12(3),* e0173636.
R15	Fuchs, C., Schreier, M., & Van Osselaer, S. M. (2015). The handmade effect: What's love got to do with it? *Journal of marketing, 79(2),* 98-110.
R16	Homans, G. C. (1958). Social behavior as exchange. American journal of sociology, 63(6), 597-606.
R17	Judge, M., Fernando, J. W., Paladino, A., & Kashima, Y. (2020). Folk theories of artifact creation: How intuitions about human labor influence the value of artifacts. *Personality and Social Psychology Review, 24(3),* 195-211.

R18	Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. Journal of economic literature, 41(4), 1188-1239.
R19	Kosinski, M. (2021). Facial recognition technology can expose political orientation from naturalistic facial images. Scientific reports, 11(1), 100.
R20	Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications, 8(1)*, 1-11.
R21	Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology, 40(1),* 91-98.
R22	Lewandowsky, S., & Pomerantsev, P. (2022). Technology and democracy: A paradox wrapped in a contradiction inside an irony. Memory, mind & media, 1, e5.
R23	Lorenz-Spreen, P., Geers, M., Pachur, T., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Boosting people's ability to detect microtargeted advertising. *Scientific Reports*, *11(1),* 15541.
R24	Mills, S., Costa, S., & Sunstein, C. R. (2023). AI, Behavioural Science, and Consumer Welfare. Journal of Consumer Policy, 46(3), 387-400.
R25	Magni, F., Park, J., & Chao, M. M. (2023). Humans as Creativity Gatekeepers: Are We Biased Against AI Creativity? *Journal of Business and Psychology*, 1-14.
R26	Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences.*
R27	Newman, G. E., & Bloom, P. (2012). Art and authenticity: the importance of originals in judgments of value. *Journal of Experimental Psychology: General, 141(3),* 558.
R28	Rasmussen, S. H. R., Ludeke, S. G., & Klemmensen, R. (2023). Using deep learning to predict ideology from facial photographs: expressions, beauty, and extra-facial information. Scientific Reports, 13(1), 5257.
R29	Reich, R. (2022). Now AI can write students' essays for them, will everyone become a cheat. *The Guardian, 28.*
R30	Shank, D. B., Stefanik, C., Stuhlsatz, C., Kacirek, K., & Belfi, A. M. (2023). AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied, 29(3),* 676.
R31	Skitka, L. J. (2012). Cross-disciplinary conversations: A psychological perspective on justice research with non-human animals. Social Justice Research, 25, 327-335.

3 Module C: Historical perspective

3.1 Introduction

3.1.1 Setting the stage

The transformative impact of technological progress on society has always generated concerns and excitement about the effects of the introduction of artificial agents to human agency and flourishing. Since the First Industrial Revolution, the enthusiasm for the opportunities offered by automation has equaled the fear of a dystopian technocratic society (Postman). In this scenario, the relationship between human users and their technological tools would be reversed if, following the economic and political interests of the few in power, the attainment of agency by machines were to translate into dominance over humans. This concern was captured early on by Isaac Asimov in his famous Three Laws of Robotics, which aimed at providing an ethical compass that is still relevant and widely quoted.[16]

The premises of this debate have fundamentally changed since the advent of the so-called Fourth Industrial Revolution (Schwab), brought up by the exceptional progress in the fields of AI and robotics since the 2010s. What distinguishes this new course from the previous Digital Age is the unprecedented level of pervasiveness and autonomy of intelligent artificial agents and systems, as well as the speed at which the technological changes in the field of AI are occurring. From recommendation engines suggesting which song to listen to, or which charitable cause to support, to industrial and military applications of big data, AI is impacting every aspect of life and reshaping the way people earn their living (e.g. gig economy, see Wood et al.), interact with each other (e.g. virtual reality platforms like the Metaverse, see Boellstorff), consume (e.g. AI used by Amazon, see West), construct and perceive their own identity (Kosinky). The long-feared prospect of a generation of intelligent machines ruling over humans is perhaps less spectacular than what we learnt to expect from science fiction, but not less likely: when an AI-operated system can rule in a court of law (Alarie et al.), deny access to a mortgage (Anderson et al.), or decide which military target to hit (de Swarte et al), there is indeed ground for concern.

Yet, there is a risk in interpreting such changes to society and democratic participation brought by the latest developments in the field of AI and big data as something unparalleled. While some aspects are indeed peculiar to the specific features and affordances of these technologies, today's core preoccupations around trust and freewill are the same posed by technological agents throughout history. For instance, the then disruptive new technologies of reading and writing were famously decried by Socrates in the Platonic dialogue *Phaedru*s as something that would have severely impaired people's ability to memorise and retain information.[17] Similarly, Swiss scientist Conrad Gessner in his book *Bibliotheca Universalis* published in Zurich

[16] The Three Rules of Robotics were theorised in the context of Science Fiction literature, as they first appeared in the story 'Runaround' (1942), later collected in the book *I, Robot* (1950). This circumstance supports the claim, crucial for the present report, that fictional narratives are extremely relevant to the history of KTs, not simply because they offer a commentary, but because they in fact shape and influence technological progress, its direction, its cultural relevance, and society's understanding and evaluation of it.

[17] In the platonic dialogue Socrates condemns Theuth, king of Egypt, considered to be the inventor of written language, of being responsible for people's loss of memory and wit and declares: "this invention will produce forgetfulness in the minds of those who learn to use it, because they will not practise their memory. Their trust in writing, produced by external characters which are no part of themselves, will discourage the use of their own memory within them. You have invented an elixir not of memory, but of reminding;

in 1545 expressed great concern over the information overload caused by the advent of the printing press, which massively increased the number of published books. The societal and political repercussions of this did not escape Gessner, who called upon kings and queens to solve the situation (Blair 2003: 11). Again, similar concerns around safety and psychological manipulation of young girls against predatory men raised today in connection with social media platform and deepfakes (Laffier and Rehman) were also debated in the late 19th-early 20th century when the telephone first entered many households and opened a channel of – often unwanted and uncontrolled – communication for young women (Marvin: 22-39). Similar preoccupations around election manipulation discussed in connection with AI and big data are to be found in several analyses on the influencing power of television (Cavgias et al.; Ragnedda and Glenn).

However, this is not to claim that AI and big data do not pose any unexpected and peculiar challenge, nor that what society faces today are simply old problems under slightly different circumstances. Instead, what Module C posits is that it is essential to explore the ever-present entanglement between technological affordances and cultural norms and values, and to identify its peculiar manifestations – historically, geographically, technologically connotated – as well as its constant traits. It is thus within this framework, which recognises the mutual shaping of culture and past and present 'knowledge technologies' – a definition discussed in section 1.2.1 –, that Module C analyses AI and big data as a specific and distinct instance of a centuries-long interaction. Tracing this history is not to downplay AI's peculiarities, but to contextualise them so as to fully understand them.

3.1.2 Goals and objectives

The historical contextualisation and the centrality of culture are the two crucial lenses through which Module C looks at the threats and opportunities posed by AI and big data to democratic and civic participation. Accordingly, Module C has two main goals:

The first one is to identify historical precedents in the way knowledge technologies have shaped the social so as to understand AI and big data as part of the long history of interactions between technological affordances and cultural norms, values, and practices. In other words, Module C is set to investigate how culture has adapted to the advent and evolution of knowledge technologies – such as written language, printing press, television, radio, etc. – but also how such technologies have been developed in response to cultural norms and changes. Module C recognises this mutual relationship as central to understanding the link between culture, technologies, and democracy. In this context, culture is intended as a complex system of practices, knowledge, and norms that every person possesses and that is indispensable in the negotiation between the individual and its society. Knowledge technologies, being expressions of culture as well as a medium for it – and far-from-neutral ones – are essential to this negotiation, which is ultimately what civic and democratic participation depends on.

The Module's second goal, strictly connected to the historical contextualisation, is to offer a definition of AI and big data as *advanced knowledge technologies* (AKTs), which would take into account the long history of the complex entanglement between culture, technology, and democracy mentioned above. Proposing a novel definition might seem to add confusion to a matter such as AI and big data that, while currently benefiting from a highly multidisciplinary discussion, is also rendered less intelligible due to single disciplines'

and you offer your pupils the appearance of wisdom, not true wisdom, for they will read many things without instruction and will therefore seem to know many things, when they are for the most part ignorant" (Plato: 563).

jargon and categorisation. However, we believe that adopting the open definition *knowledge technologies* and applying it to AI and big data can lead to a soberer assessment of their uniqueness due to the historical contextualisation proposed in Module C. The hope is that, by focussing on constant traits and similarities across time, our analysis will stay current beyond the present moment, since the pace and trajectory of AI development is extremely fast and unpredictable and thus makes its assessment quite volatile.

Before embarking upon the comparative analysis of AI and big data, alongside historical knowledge technologies, which will be Module C's next step, it is now essential to clarify three foundational considerations.

First, we need to explore the distinction between *knowledge technologies* and other definitions such as information technologies and understand why our focus is on the former. This exploration involves delving into the nuanced differences between knowledge and information, justifying the prioritisation of the former.

Second, the importance of adopting a historical perspective needs examination, revealing the motivations behind situating AI and big data within the extensive history of knowledge technologies. This exploration of historical context acts as a lens to uncover the evolution, trends, and paradigm shifts in knowledge technologies, enhancing our understanding of the contemporary landscape.

Lastly, the centrality of cultural processes in investigating AI and big data becomes a crucial theme, leading to an exploration of the reasons for emphasising cultural dimensions in this inquiry. This focus on cultural processes highlights the socio-cultural influences that shape the development, implementation, and impact of AI and big data.

These three crucial aspects are discussed in the following literature review, providing an overview of relevant scholarly discourse, and offering insights into the complexities of each dimension. This analysis sets the stage for a comprehensive understanding of the interconnected realms of AI, big data, and historical knowledge technologies, which is the final goal of Module C.

3.2 Literature Review and Rationale

3.2.1 Why 'knowledge' over 'information'?

3.2.1.1 Limitations of the existing definitions of 'knowledge technologies'

The term 'knowledge technologies' has not been used extensively (24,800 results on Google Scholar compared to 1,670,000 results for 'information technologies') and definitely not in a critical way, but more as an operational definition. The term was a more popular definition between 2000s-early 2010s and it was often used to indicate practical tools (often software) for knowledge management (Garavelli et al.), or to talk about the Semantic Web (Rigau et al.). This means that knowledge technologies are usually intended solely as digital and computer technology (Milton: 13).

The definition is also found in publications and projects written by scholars who are not native speakers of English (many in the Balkans and Eastern European countries, and Italy) and whose main audiences are not Anglophone academics, used probably because it provides a more literal and thus accurate translation. More recently, the label 'knowledge technologies' has been used to designate educational tools – most exclusively digital ones – making remote learning possible during the COVID-19 pandemic. In these cases (Stewart and Khan; Dionisio-Flores et al.) the word 'knowledge' of the definition stands for 'knowledge acquisition' and has a specific pedagogic connotation.

This literature often considers how knowledge – understood as content – could be successfully transferred and managed by means of employing dedicated ‘knowledge technologies.’ More importantly, many of these analyses, especially the one developed within the field of knowledge management, tend to adopt the definition as self-explanatory. This is because they mostly focus on establishing what ‘knowledge’ is and, once satisfied with a definition, they assume that every tool used to share and manage it, is by necessity a ‘knowledge technology’.

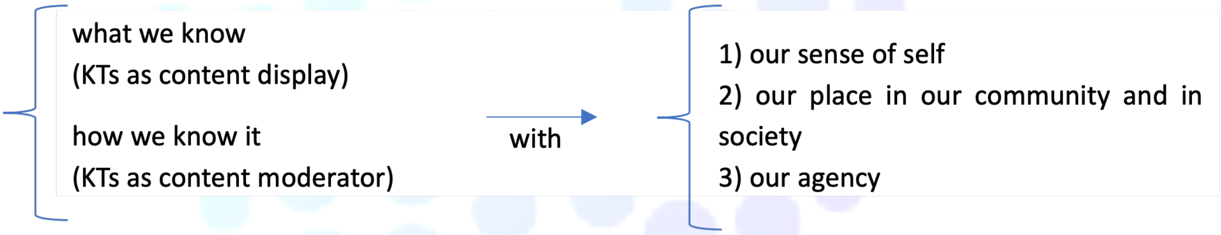
3.2.1.2 Difference between information and knowledge

In the pursuit of delineating the essence of knowledge technologies within the Knowledge Technologies for Democracy (KT4D) framework, a critical point of consideration is the distinction between knowledge technologies and the more prevalent realm of information technologies. A noteworthy perspective emanates from the edited volume titled *Information Technology for Knowledge Management* (Borghoff and Pareschi). While belonging to the scholarship interested in information management discussed in the previous section, it nonetheless provides an interesting distinction between *information* and *knowledge* which aligns with our project's objectives and that it is worth quoting in its entirety:

> Knowledge is quite different from information, and managing knowledge is therefore decisively and qualitatively different from managing information. **Information is converted into knowledge through a social, human process of shared understanding and sense-making at both personal level and organizational level.** Managing knowledge starts with **stressing the importance of people**, their work practices, and their work culture, **before deciding whether or how technology should be brought into the picture**. Information management, on the other hand, often starts with a technological solution first – with consideration of people’s work practices and work culture usually a distant second (Holtshouse: V).

When extrapolated beyond the analysis of knowledge technologies for work practices and environments, these considerations encapsulate the intricate interplay between society, culture, and technology that underpins the analysis develop in Module C. The deliberate shift from a focus on information (the substance of knowledge) to knowledge itself (the process of sense-making) facilitates an engagement with a more culturally and socially intricate conceptualisation of technologies. This conceptualisation aligns harmoniously with the broader discourse in Media Studies (see section 1.2.3.1), underscoring the significance of a cultural and historical lens in comprehending the multifaceted dynamics of knowledge technologies.

Thus, when using the definition *knowledge technologies*, we aimed at addressing the link between:



what we know
(KTs as content display)

how we know it
(KTs as content moderator)

with

1) our sense of self
2) our place in our community and in society
3) our agency

Only when these technologies are used to gate the information individuals have access to, to restrict the messages shaping their perception of public opinion, and to control their choices and interactions, we start to see the space of both opportunity and risk opening for these technologies to enhance or harm democracy, and this is the space of *knowledge technologies*.

3.2.1.3 Defining knowledge in the age of self-learning AI

The tendency of focusing on content rather than on processes encountered in the existing literature adopting the definition of KT discussed above, risks imposing a simplistic definition of what knowledge technologies are. This is because it assumes that the mediation operated by the technology is a neutral one, as the tool is simply a carrier and an organiser for content and does not affect its nature. This is even more problematic when we consider the paradigmatic change that self-learning AI has brought about in recent years. In the case of self-learning agents, knowledge is not simply a content that needs to be managed and organised, as these systems do not merely support humans' attainment of knowledge, but indeed replace them in the process of acquisition.

While this change might be upsetting, as knowledge creation and acquisition have been traditionally considered human prerogatives, this situation might also offer an opportunity to explore with a renewed critical awareness the until now under-researched link between 'knowledge' and 'technology'. This is because artificial neural network technologies require that we operationalise a complex concept like 'knowledge' in order to teach an artificial agent to reproduce what we humans have historically accomplished in an assisted but yet autonomous way. The dominant approaches in this field look at AI knowledge from the perspective of logics, computational knowledge, linguistics, and semantics (Aamodt and Nygard; Guarino; Levesque; Zhuang et al.). This approach risks missing the opportunity of recognising AI as a sandbox for exploring the role of culture in shaping the link between knowledge and technology. Many ever-present cultural patterns – of dominance, discrimination, manipulation, but also of inclusivity and reparation – are automatised and played out in front of us when it comes to knowledge produced and managed by AI agents. What used to be tacit is brought to light, and this can lead to positive action and change.

Also, current definitions of knowledge used in describing self-learning agents (Koggalahewa et al.; Stein et al.) and developed within the fields of logic in computer science and cognitive science risk adopting a universalistic approach that erases the cultural peculiarities which are instead central to the way people understand and interact with both knowledge and AI technologies. Alan F. Blackwell, Addisu Damena and Tesfa Tegegne, in a recent article dedicated to the peculiar approach to AI research in Ethiopia, challenge the claim that current developments in AI technologies are "concerned with understanding of humans" (370), as if 'to be human' were a universally shared condition that applies to all people in the same way. Consequently, they reject the idea that "the fundamental understanding of humans [is] necessarily universal" (370) and assert that the skills and behaviours which are considered to be 'human-like' in AI – knowledge acquisition included – are instead Western-centric interpretations of otherwise multifaceted and culturally defined concepts. They thus provocatively ask:

> Will such understanding be the same wherever it is investigated, regardless of who the humans are, or of what culture they have inherited, or what their economic and political circumstances might be? Such attempted universalism seems extremely unwise, despite the AI reliance on supposedly universal principles of cognitive science (critiqued rather comprehensively by Geoffrey Lloyd in his book *Cognitive Variations* (Lloyd 2007). (Blackwell, Damena, Tegegnepp: 370-371).

It is evident that, while the development in AI technologies offers the chance for new critical approaches to the understanding of what knowledge technologies are, there is also a concrete risk that the old approach dominating the field of knowledge management in the early 2000s-2010s will now be replaced with another one that, while recognising the difference between 'information' and 'knowledge', will still ignore the pivotal role of culture.

In adopting the definition of 'knowledge technologies' in an expansive way, the goal of this Module within the KT4D project is to highlight how these cultural patterns are essential in understanding the link between 'knowledge' and 'technologies' and how they can only be understood when adopting a historical approach. The rich and complex past analyses of KTs – although named differently – can contribute to the discussion by adding *culture* (intended as situated practices and knowledge across space, time, and technologies) to a discourse that is often solely preoccupied with technical aspects and with the present time.

3.2.2 Why the historical perspective?

3.2.2.1 Challenging the concept of newness

The rationale for adopting a historical perspective in the analysis of AI and big data derives from one of the central hypotheses underpinning the KT4D project: that in order to understand the social and cultural impact of these advance knowledge technologies and their impact on civic and democratic participation, we need to contextualise them within the long history of interactions between technological affordances (writing, printing, television, etc.) and cultural norms, values, and practices. By doing so, Module C aims at providing a novel historical perspective allowing for a more sober and critical engagement with AI technologies, whose novelty and impact are often overhyped and thus misunderstood. It is only through a historical examination that significant precedents and paradigms can be fruitfully examined and tested against modern challenges.

The conviction that a full understanding and critique of the current impact of AI and big data on democratic participation cannot prescind from a historical contextualisation stems from the decades-long debate in the field of Media Studies on the concept of 'newness' applied to media and information-communication technologies. The need to critically unpack this concept came with the mass diffusion of the label 'new media' in the mid-1990s, when the definition was applied to digital media and web-related communication technologies. Influential media scholars like Friedrich Kittler (1997)[18] and Lev Manovich (2002 and 2003) interpreted the advent of modern computer technologies as a moment of rupture from the past and thus adopted the definition of 'new media' to mark the beginning of a new era in the way people create and share knowledge and information.

In response to these analyses and to the general enthusiasm towards the Web and digital media, other scholars around those same years started challenging the very concept of 'newness' as the result of a calculated hype serving the interest of tech companies, or, as in the case of many Media Studies scholars, of an excessive focus on the technical aspects to the detriment of the cultural and social dimension of media technologies.

One of the first and still authoritative sources is Carolyn Marvin's book titled *When Old Technologies were new* (1990), in which she analysed two 'new media' of the 19th century: the electric lights – intended as a medium in sense indicated by Marshall McLuhan (1964: 8-9 and 52), and the telephone. In focusing on these old technologies, Marvin did not simply aim to demonstrate how every invention was once new, but instead focussed the attention on how the very concept of novelty is culturally and socially determined and, in turn, how any new media imposes and shapes social norms and hierarchies. In the introduction to her book, she immediately made clear that "the early history of electric media is less the evolution of technical efficiencies

[18] Kittler wrote that "The last historical act of writing may well have been the moment when, in the early seventies, Intel engineers laid out some dozen square meters of blueprint paper […] in order to design the hardware architecture of their first integrated microprocessor" (147).

Funded by the European Union

KT4D has received funding from the EU's Horizon Europe research and innovation programme under Grant Agreement no. 101094302.

in communication than a series of arenas for negotiating issues crucial to the conduct of social life; among them, who is inside and outside, who may speak, who may not, and who has authority and may be believed" (Marvin: 4).

Therefore, what matters is not the technical aspects, the nature of the new media – which is what interested scholars like Kittler and Manovich – but the social and cultural substratum that receives and makes sense of the new technologies. Evidently, the changes to such substratum do not happen abruptly and are not determined solely by technological progress. To further prove this point, Marvin further remarked the focus of her research and consequently of her book was

> shifted from the instrument to the drama in which existing groups perpetually negotiate power, authority, representation, and knowledge with whatever resources are available. New media intrude on these negotiations by providing new platforms on which old groups confront one another. Old habits of transacting between groups are projected onto new technologies that alter, or seem to alter, critical social distances. […] Old practices are then painfully revised, and group habits are reformed. New practices do not so much flow directly from technologies that inspire them as they are improvised out of old practices that no longer work in new settings (Marvin: 5).

Marvin's emphasis on social and cultural practices and on their resistance to technological changes is an essential aspect in her understanding of new media. This approach is in line with what this Module aims at demonstrating, meaning that AI and big data should be inscribed and understood as part of the long history of knowledge technologies. Like reading and writing (and the long line of knowledge technologies that followed them, from the printing press to television and the Internet), also AI and big data institute and threaten established hierarchies and disrupt interactions between members of a community, for instance via data analytics, algorithmic filtering of information sources and microtargeting. These examples all represent disruptions in our relationship with the way we apprehend the world, narrativise the reality we see, and act upon these interpretations so as to maximise the quality of our lives. However, these disruptions are not peculiar to AI and big data. Instead, a history of these issues and of how people adapted to and dealt with them can be traced, like Marvin did, to past examples.

Lisa Gitelman (2006) offered criticism of the concept of 'new media' as an ontological reality similar to the one raised by Marvin. Gitelman too analysed two case studies, one old medium and one new, at least at the time of the publication of her book: the phonograph and the World Wide Web. Beside the focus on the social and cultural nature of mediation and, thus, of communication and information technology itself, she offered an important remark on the permanency and resilience of cultural norms and processes in the face of rapid technological change. Gitelman wrote:

> The introduction of new media […] is never entirely revolutionary: new media are less points of epistemic rupture than they are socially embedded sites for the ongoing negotiation of meaning as such. Comparing and contrasting new media thus stand to offer a view of negotiability in itself – a view, that is, of the contested relations of force that determine the pathways by which new media may eventually become old hat (Gitelman: 6).

The need for a comparative approach that alone can reveal the process of cultural negotiations that media technology enact is in line with the approach that our own analysis adopts and that ultimately justifies our chosen historical perspective. One of the overarching research questions of the KT4D project asks how we can place enhanced cultural processes, by their very nature subtle and intangible, at the heart of an investigation of technology. Gitelman's definition of new media as "sites for the ongoing negotiation of meaning" thus suggests a valuable starting point for our investigation.

One last important contribution to the critical investigation of the concept of newness in media that we ought to consider is the concept of 'remediation' famously theorised by Bolter and Grusin. This posits the constant and mutual shaping of old and new media and consequently establishes the impossibility to consider any communication technology in isolation. In open disagreement with scholars supporting the idea of an unprecedented change in the media panorama of the late 1990s due to the advent of the Web – not much different from what is happening today in relation to AI – Bolter and Grusin argued that "No medium today, and certainly no single media event, seems to do its cultural work in isolation from other social and economic forces. What is new about new media comes from the particular ways in which they refashion older media and the ways in which older media refashion themselves to answer the challenges of new media" (15).

In this case, differently from what happens in Marvin's and Gitelman's analyses, the focus is firmly on the technologies rather than on socio-cultural processes. Nonetheless what is relevant to Module C's comparative analysis of past and present knowledge technologies is that the concept of remediation postulates the need for a contextual analysis, both synchronic and diachronic. This in turn supports the claim that old media and AI and big data are not to be understood as sequential steps in the evolution of knowledge technologies, one replacing the next one by rendering it obsolete, but instead should be regarded as part of a complex system that needs to be analysed in its entirety.

It is important to point out that these scholars challenged the concept of newness in relation to 'media', while in this Module, and in the KT4D project more in general, we choose to focus on 'knowledge technologies', a label that, while we are in the process of defining it (see section 1.2.1), it is nonetheless a non-negotiable point of reference. Media, with its accent on communication, speaks of a necessarily public dimension, because, even if consumed in solitude, any medium implies a broadcaster or a sender, and an infrastructure. Our definition, with its accent on 'knowledge', encompasses both the individual and the social dimension of sense-making. Moreover, it considers not only the process of mediation and acquisition of knowledge, but also the preceding and following steps, meaning the precondition that makes the acquisition of knowledge possible, desirable, or needed, and the consequences of such acquisition in terms of agency, freedom, and awareness. Differences notwithstanding, the focus on the socio-cultural dimension of information and communication technologies discussed in the field of Media Studies (further discussed in section 1.2.3) is an approach that Module C will heavily borrow and apply in its analysis of past and present KTs.

3.2.2.2 Using the past to understand AI and big data

As famously stated by Howard Rheingold (1985), the pioneering theorist of Internet technologies and virtualisation, it is impossible to understand where mind-amplifying technology is going unless we understand where it came from. However, there is one aspect in need of clarification before adopting an approach that centres such historical contextualisation and this is the question of scale and nature that supposedly distinguish past and advanced knowledge technologies. In order to meaningfully compare the impact that 'old' knowledge technologies had on civic and democratic participation with the one imposed by AI and big data, one must first assert that the changes brought about the new technologies – or at least the ones salient for our analysis – are different only in scale, but not necessarily in nature.

That this should be the case is supported by the growing number of academic analyses reading AI and big data not just against past technologies and systems, but as direct evolution of what preceded them. Like in the case with the definition of 'new media' challenged by scholars wary of the hype around the Web in the late 1990s and early 2000s, today's scholars who are interested in historically contextualising AI do so to counter the claims, coming from tech companies and mainstream media alike, overstating the unprecedented revolution brought by these technologies. The current effort to historically contextualise AI and big data looks at four main aspects.

First, there is a growing interest in the analysis of the power structures and social hierarchies that have allowed the recent rise and spread of modern AI technologies and systems. One main point of discussion is the role of historical colonialism (Adams; Hao) in creating the premises for current forms of exploitation and data extraction in part of the world that, while formally emancipated from foreign domination, are still subjected to Western economic power and political influence, which are often further exerted with the aid of AI and big data. The claim advanced is that it would be impossible to assess the impact on democratic participation of these technologies without understanding the socio-political context that make them viable and that, in turn, they reinforce.

Second, attention has been devoted to the technical and material aspects of AI which are borrowed and inherited from previous technologies. This approach is in line with the previously discussed concept or 'remediation' and aims at highlighting how affordances and constraints of past technologies that AI improves upon, necessarily shape and influence it. This is the case, for instance, of the analysis of the dependency of deep learning machine vision from traditional photography conducted by Daniel Chávez Heras and Tobias Blanke. They demonstrate how machine vision inherited from photography "its technical regimes and epistemic advantages" (1153) so what is labelled and detected by algorithms is not 'the world', but a (culturally and socially determined) vision of the world that two centuries of photography has previously codified. Their claim is thus that computer vision should treat "photographs not as detections of the world, but as measurements of these beliefs" (1158). To fully understand these beliefs, they posit, we ought to consider the history of photography from which machine vision stems. Therefore, investigating where AI and big data come from is essential not simply to critically understand their cognitive and cultural impact, but also to change their value system and to redirect their purposes.

Third, consideration has been given to the cognitive impact of AI-generated content and to the consequent moral panic that this is ensuing by comparing the present situation to past instances. This is, for example, the case with the study currently undertaken by Joshua Habgood-Coote, researcher at the University of Leeds in Philosophy of Language, who investigates the threats posed to our epistemic practices by deepfake videos. In a recent article, Habgood-Coote claims that both people's current lack of trust in the images we see, due to the proliferation of deepfake and AI-generated content, and the consequent need to develop knowledge and cognitive tools in response to such changes, are not at all unprecedented. Instead, as he documents in his analysis, there is a long history of photographic manipulation that constitutes an important precedent, like in the case of the "composograph". This was a forerunner method of photo manipulation and is a retouched photographic collage popularised in the 1920s by American publisher Bernarr Macfadden who used to produce fake sensationalist pictures of celebrities. The analysis interestingly focuses on people's reaction to this fraud and on the cultural and cognitive tools and strategies developed in reaction to it. Far from underplaying the issues raised by AI-generated content, studies like this one recognise people's agency and awareness when confronted with unreliable sources and identify virtuous processes of shared meaning-making from which to learn.

The fourth and final aspect concerns the need for a new theoretical approach to modern AI that looks back at past conceptualisations of intelligent systems and autonomous agents. In the last few years some scholars have advocated for a 'return to cybernetics' (Bell et al; Pangaro; Pickering 2010), intended as the highly interdisciplinary and human-centred approach to human-computer interaction and intelligent systems laid out in the early 1950s and went out of fashion in the late 1970s. To the proponents of this approach, historical cybernetics, within which the research on AI originated, could offer a valid alternative to the current trends. First, cybernetics aimed at offering a general epistemology that encompassed but also exceeded the technical issues at hand and thus provided a holistic and farsighted approach (Johnston; Vidales). Second, cybernetics was a truly interdisciplinary and collaborative field to the point of being described as 'anti-disciplinary' (Pickering 2013). Lastly, due to the lack of direct applicability of many of its projects and inventions (Pangaro: 17), and to the disastrous outcome of technological applications during WWII, to which cyberneticians contributed (Galison), endeavoured to follow strong ethical principles (Wiener). Therefore, in the hopes of the promoters of a return to cybernetics, its holistic approach would remedy the present utilitarian, task-driven vision of AI and allow for a complex, humanistic one; its 'anti-disciplinary' attitude will respond to the call for interdisciplinarity, in and outside of academia, in relation to the study of complex systems such as, for instance, the Internet of things (Adamson et al.); its relative autonomy from invested interests would set an example for a more ethical approach to AI, currently dominated by economic and military goals, when not even by antidemocratic and manipulative forces.

Looking at past instances of knowledge technologies thus allows to contextualise and fully understand the hierarchies of power and domination, the technical and aesthetic beliefs and assumptions, the cognitive impact and literacy strategies, and the epistemic system upon which AI and big data rely. It is from this growing scholarship that looks at the past in order to understand the present that the historical contextualisation adopted in Module C stems from.

3.2.3 Why the cultural perspective?

3.2.3.1 Knowledge technologies as systems

Technological changes, at least when we consider specific inventions and manufacts, occur at a fast pace and, when it comes to AI and big data, such changes are happening at an even higher speed than the one ever witnessed before. The call is thus for societies to adapt their cultural responses to these new technologies in order to master these tools, and guide and regulate their implementation to avoid being manipulated and overwhelmed. However, if we assume that technological progress does not exist outside culture, we must also concede that our culture has already changed in order for these advancements to even happen.

Determining the direction of this transformation – from culture to technology or vice versa – is what the long debate between the supporters of technological and cultural determinism has always tried to discern: is it technology that imposes cultural changes, or is it culture that makes any technological advancement possible? It is worth considering the positions famously held by two of the most renown exponents of the two fronts, Marshall McLuhan, advocating for technological determinism, and Raymond William for social determinism. In his 1962 book, *The Gutenberg Galaxy,* dedicated to the technology of writing McLuhan stated that its invention and evolution marked all major steps in human history. He wrote that:

> Any technology tends to create a new human environment. Script and papyrus created the social environment we think of in connection with the empires of the ancient world. […] Technological environments are not merely passive containers of people but are active processes that reshape people and other technologies alike. […] Printing from movable types created a quite unexpected new environment – it created the public. Manuscript technology did not have the intensity or power of extension necessary to create publics on a national scale. What we have called "nations" in recent centuries did not, and could not, precede the advent of electric circuitry with its power of totally involving all people in all other people. (McLuhan 1962: XXVII).

Opposite convictions were held by William who in his 1971 book *Television: Technology and Cultural Form* asserted that any new technology, such as the printing press, is always developed in response to specific social, political, and cultural changes rather than these transformations proceeding from the introduction of the new technology:

> The development of the press […] was at once a response to the development of an extended social, economic and political system and a response to crisis within that system. […] In Britain the development of the press went through its major formative stages in periods of crisis: the Civil War and Commonwealth, when the newspaper form was defined; the Industrial Revolution, when new forms of popular journalism were successively established; the major wars of twentieth century, when the newspaper became a universal social form. […] What matters, in each stage, is that a technology is always, in a full sense, social. It is necessarily in complex and variable connection with other social relations and institutions, although a particular and isolated technical invention can be seen, and temporarily interpreted, as if it were autonomous (14).

In Module C we do not espouse either of these positions exclusively, but rather, following a well-established and today dominant tendency, we combine and take advantage of the insights offered by both as we understand them to not be mutually exclusive. This is because, rather than seeing culture and technologies as two self-defined forces in opposition, we consider knowledge technologies as complex systems made of cultural, social, and technical components that constantly and mutually shape each other in a process that has no direction and can thus be apprehended only as a whole. Indeed, we use the term 'advanced knowledge technologies' to refer to the assemblages of advanced processing and big data, not according to the kinds of methods that are used to develop them, but rather to those specific implementations or these technologies that are most likely to disrupt civic participation and democratic processes by intervening in the manner in which individuals develop their sense of themselves, others, and the world around them. What we aim to avoid is thus an essentialist and limiting definition of what knowledge technologies are and instead understanding them as *systems*, which is in line with the dominant definitions of technologies developed within the field of Media Studies.

Donald Mackenzie and Judy Wajcman, for instance, offer a three-level definition of technology. First, they define technologies as sets of physical objects, though they also concede that "few authors are content with such a narrow 'hardware' definition" (3). Second, they define the concept as referring to all the human activities associated with a particular technology, either those directly linked to a particular machine (e.g., the programming work essential to make a computer function) or the social behaviours a technology prescribes (for instance, urban habits developed in response to mass motorization). Finally, Mackenzie and Wajcman consider technologies as forms of knowledge, meaning the practical and theoretical know-how necessary to design, repair, and operate machines.

Similarly, Ursula Franklin, elaborating on Jacques Ellul's concept of technique (1954), describes technology as practice and rejects any definition that limits it to the material: "[t]echnology is not the sum of the artifacts, of the wheels and gears, of the rails and electronic transmitters" (10) Technology is a system. It entails far more than its individual material components. "Technology involves organisation, procedures, symbols, new words, equations, and, most of all, a mindset" (Franklin: 10).

Also, Gitelman, in her previously mentioned analysis, provides a definition of media in line with the one we propose of knowledge technologies that stresses the same complex entanglement of technical and cultural aspects and how this convergence must be understood in its complexity. She writes:

> I define media as socially realized structures of communication, where structures include both technological forms and their associated protocols, and where communication is a cultural practice, a ritualized collocation of different people on the same mental map, sharing or engaged with popular ontologies of representation. As such, media are unique and complicated historical subjects. Their histories must be social and cultural, not the history of how one technology leads to another, or of isolated geniuses working their magic on the world. (Gitelman: 7)

Finally, another important input comes from what Kember and Zylinska call a performative approach to mediation. Again, although their analysis focuses on 'media' rather than 'knowledge technologies', it is possible to extrapolate relevant points for our analysis in light of the shared attention to the cultural aspects and historical approach. Kember and Zylinska apply the concept of performativity to the understanding of information and communication technologies and posit that "media are generative, that is, that they are part of the material world and do not thus exist apart from it. Neither a reflection of nor a mask for the social, media actively contribute to the production of the social. In other words, media perform the social – sometimes alongside and sometimes in conflict with other agencies that are not solely establishment or anti establishment" (38). This position evidently builds upon Bruno Latour's and Michel Callon's "Actor Network theory," which famously challenges the distinction between linguistic, social, technological, and natural realms, a distinction on which traditional sociological studies are predicated.[19] Indeed Kember and Zylinska write of mediation as a "multiagencial force that incorporates humans and machines, technologies and users, in an ongoing process of becoming-with that is neither revealed nor concealed but rather apprehended intuitively – inevitably from inside the process" (40).

Module C, following in the steps of this scholarship, recognises that both threats and opportunities pose to democratic participation by AI and big data – and by any kind of KTs more in general – arise from this everlasting negotiation, in which established cultural values and norms are not passively shaped by technological progress, nor actively determining its course. Human culture is not an endangered territory, nor a *post hoc* cure to unethical applications of AI, but one among the active forces implicated in the process and it needs to be recognised and studied as such.

3.2.3.2 The cultural dimension of AI ethics

One of the goals of the KT4D project, and of this Module specifically, is to investigate the cultural dimensions of ethical AI, understandood in terms of languages and discourses, national or regional identities, religions, beliefs and practices, values and tolerances, etc. These elements are often disregarded by traditional approaches to AI ethics that instead focuses on more universal and abstract values.

However, when we consider the major threats and downfalls of AI systems in relation to democratic and civic participation, we notice that they tend to happen whenever these technologies – developed as standardised

[19] See: Callon, Rip, and Law; and Latour.

and neutral tools and marketed globally as such – impinge on the cultural values and social structures of the communities that adopt them. For this reason, Module C will consider more and lesser known case studies that demonstrate the need for an approach to ethical AI that considers its cultural dimension. It will do so by focusing specifically on:

• *The complexity and heterogeneity of identities*, which we address by adopting the Feminist analytical framework of intersectionality (Crenshaw). People manage different aspects of their identity in different contexts and respond to situations differently depending on the social role they are playing at the time. Knowledge technologies, including AI and big data, can either provide tools empowering people to express their complex and stratified identities, or can enforce patterns of discrimination, which are further crystallised due to the technology's affordances. For instance, it has been proven, especially during the Covid-19 pandemic (Leslie et al.) that AI systems used in the medical sector are trained on datasets that reflect the differences in treatment that white patients and patients of colour receive. Those differences are immortalised in data, which are then used to train algorithms that ultimately perpetuate the discrimination.

• *The importance of cultural-difference awareness*, which we draw from Geert Hofstede's Cultural Dimensions Theory as our first and general point of reference.[20] Hoftede's framework identifies six key dimensions (power distance, uncertainty avoidance, individualism-collectivism, masculinity-femininity, short vs. long-term orientation) aimed at capturing cultural differences across countries. While scholars have pointed out many limitations inherent to this framework (e.g. the focus on the nations as homogeneous cultural sites (see McSweeney), and the lack of women's perspective (see: Moussetes), its usefulness resides in its general statement against the claim that digital technologies are erasing cultural differences. Hofstede's framework challenged the theory of the 'global village' and demonstrated the local dimension of culture. Indeed, while software product releases tend to be international, their use and applications depend on local habits, norms, and communities. An example is offered by the fast and positive reception of cryptocurrency in the Islamic world due to its compliance with Islamic banking that prohibits usury and speculation and thus any form of investments (Khan and Rabbani). In recent years some Islamic scholars deemed cryptocurrencies halal and thus religiously permissible and are trying to prove that rules and regulations from sharia are fully compatible with digital blockchain technology. Religious beliefs are then what made the new technologies acceptable and indeed desirable.

• *The importance of people's values in technology adoption*. While this is a virtuous principle that guides the well-established field of User Experience (UX) Design, it is also true that its applicability often depends on designers and programmers who have, by training, limited knowledge of cross-cultural issues (Lachner et al.). It is a recurrent experience for people to have wrong expectations about software and technologies and misuse them with more or less severe consequences, or to deliberately choose a different purpose for their tool. This is the case, for instance, with a growing number of parents using Apple's AirTags to track their children and ensure their safety. When Apple released AirTags in 2021, the company clearly stated that they were not to be used for children or pets, only on inanimate objects, but parents and caregivers are choosing to do otherwise. It would be easy to dismiss this as a reckless decision that speaks of their technological

[20] Cultural Dimensions Theory, developed by Geert Hofstede, is a framework used to understand the differences in culture across countries. Hofstede's initial six key dimensions include power distance, uncertainty avoidance, individualism-collectivism, masculinity-femininity, and short vs. long-term orientation. Later, researchers added restraint vs. indulgence to this list. The extent to which individual countries share key dimensions depends on a number of factors, such as shared language and geographical location.

Funded by the European Union
KT4D has received funding from the EU's Horizon Europe research and innovation programme under Grant Agreement no. 101094302.

illiteracy. However, some newspaper articles and journal investigations (Kelly; Greenaway) uncovered a more complex picture of why parents, in negotiating with their kids the boundaries of freedom and autonomy, recur to AirTags: in a society in which technology poses new threats to young children (e.g. online grooming), it is only logical that parents also look for technological remedies.

To test our hypothesis, we have presented the three case studies mentioned above to the participants of the first workshop for our Use Case 4 (see deliverable 1.2), which invited software developers to assess and discuss their approaches to ethical AI. The cultural dimensions of the three issues raised by AI technologies were deemed by the participants the most elusive and difficult to deal with when designing AI systems and software, and the one for which a comprehensive and clear understanding is missing. This has reinforced our conviction that to focus our analysis of past and present knowledge technologies on the entanglement between cultural and technological aspects – in line with the scholarship discussed in the previous section – is a much-needed contribution that our project can offer.

3.3 Methodology

3.3.1 Original contribution

As it has emerged from the literature review, many scholars from different fields have discussed the benefit of providing a historical contextualisation to the analysis of the impact of AI and big data, as well as the need to recognise the central role of culture in the development and adoption of new technologies. While Module C takes advantage and builds upon this vast and relevant scholarship, it will also provide an original contribution to this debate. This is due to the two particular approaches that distinguishes the KT4D project from similar existing academic endeavours:

- First, the project's focus on *knowledge technologies* and the importance, discussed in the previous sections, of this open and novel definition that identifies a category distinct from the more commonly used ones of media and information technologies. This is because the project recognises the process of knowledge creation and sharing as being central to democratic participation.
- Second, this Module adopts a systematic approach to the comparative analysis of past and present knowledge technologies from a cultural studies perspective. As described in the previous sections, there is a growing number of scholars historically contextualising AI tools and systems and drawing parallels with past technologies. However, their focus is usually limited to one issue or technological application at a time, e.g., in the already mentioned comparison between deepfakes and photographic manipulation of the 1920s. What Module C aims to accomplish, instead, is a comprehensive map of past and present knowledge technologies so as to identify general trends, divergences, and similarities. The overarching theme threading these case studies together is the definition of *knowledge technologies* to which they all refer to.

3.3.2 Establishing precedents: not a list of technologies, but a list of issues

To adopt a definition of technologies as systems (see section 1.2.3.1), and of knowledge technologies as systems specifically involved in the process of sense-making, it means that past and present examples of KTs can only be understood historically and contextually. For this reason, the analysis developed in Module C will not consider a list of specific examples of past KTs (e.g. the printing press, television, Web 1.0) and then compare them one by one to advanced KTs (AI and big data). This is because an approach of this sort would assume that the technological element is preponderant compared to the human one (something that is closer to the definition of *information* rather than *knowledge* technologies, see section 1.2.1.2), a hypothesis that many scholars has refuted, as shown in the literature review, and to which Module C subscribes.

Furthermore, this approach would lead to erroneously consider KTs as tools rather than systems and would impose an abstract and essentialist idea of what each technology is, which is something that Gitelman, among others, criticises and warns against:

> So it is as much of a mistake to write broadly of 'the telephone', 'the camera', or 'the computer' as it is 'the media', and of – now, somehow, 'the Internet' and 'the Web' – naturalizing or essentializing technologies as they were unchanging, 'immutable objects with given, self-defining properties' around which changes swirl, and to or from which history proceeds. Instead, it is better to specify telephones in 1890 in the rural United States, broadcast telephones in Budapest in the 1920s […]. Specify is key. Rather than static, blunt, and unchanging technologies, every medium involves a 'sequence of displacement and obsolescences, part of the delirious operations of modernization', as Jonathan Crary puts it. (8)

Following Gitelman's recommendation, Module C will develop its comparative analysis of past and present KTs by assuming that something such as 'television' or as 'chatbots' does not exist, instead there are only historically and culturally realised interactions between people and versions of these technologies. Accordingly, the study will focus on similar issues across time and across technologies that are relevant to contextualise AI and big data (see Section 1.4.1). Indeed, to identify relevant and reusable patterns from past interaction with KTs (either as cautionary tales or as virtuous examples), Module C will first identify people's needs, fears, hopes, problems around AI and big data, drawing from the insights offered in Module A and B, and then look for similar entanglements in past interactions with KTs, without incurring in misleading generalisations such as that 'AI is the new printing press'.

The initial framework adopted in Module C to map suitable case studies considers three main categories: *agency*, *creativity*, and *identity*, which subsume the two aspects of *free will* and *trust* discussed in the analyses of Module A and B:

- *Agency* pertaining to the process of *knowledge access and sharing*. This includes issues concerning consciousness, intentionality, free will, and autonomy. This category addresses how KTs have always been used to manipulate people through propaganda and social control, but, at the same time, they have also been used to democratise access to information and to support liberation movements;
- *Creativity* pertaining to the process of *knowledge creation*. This deals with two opposing views of KTs as capable of threatening people's capacity for creativity, understood as a quintessentially human trait, and thus limiting their freedom of expression, or, at the opposite, as tools relieving people from boring menial tasks, or even offering opportunities to further enhance their creativity;
- *Identity* pertaining to the process of *knowledge acquisition*. This category focuses on the link between KTs epistemology and people's understanding of their role and place within their community. As KTs can lead to a more truthful, effective expression of one's identity and thoughts, those same technologies may pose a threat as they can enforce stereotypes, identity-based discrimination, or simply disrupt social hierarchies and cultural norms essential to people's identity building.

As it emerges from this framework, the analysis developed in Module C will not only consider the threats posed by AI and big data, but will also examine the opportunities that these technologies present and will do so by identifying historical precedents of how people have leveraged the power of KTs for good. In doing so, it is worth considering Mike Ananny's interpretation of the double role and function of algorithms used in machine learning and AI systems. Ananny states that “Algorithms are both ‘traps’ that sequester people in particular cultural worldviews, and ‘societies’ that transform how ‘people interact, associate, and think.’ They simultaneously give people options for what to do, and signal what people are expected to do and what most people do” (6). Understanding not only algorithms but KTs in general as both ‘traps’ and ‘societies’ helps us by recognising how these tools and systems have the power to hamper democratic participation and personal realisation, but can also be enablers of positive change and serve the needs of society as a whole. Indeed, what Ananny writes about algorithms “creating descriptions of the world that people use to reflect upon their identities, communicate with others, and create public life” (6), also suits our definition of KTs intended as cultural systems for shared meaning-making.

3.3.3 Chapters outline

Following the framework and principles outlined in the previous section, the analysis proposed in Module C will be organised in subsections. Each section will first identify an area of interest and a connected set of issues (problems and opportunities) that AI and big data pose today for democratic participation, and then identify historical precedents of the same issues (or at least similar) that people had to confront when dealing with past examples of KTs. The sections will be organised as follows:

1. Knowledge technologies and power structures

Threats: Asymmetries of power
Opportunities: Subversion, empowerment, and civil disobedience

2. Knowledge technologies and access to information

Threats: Disinformation and Manipulation
Opportunities: Democratising access to knowledge and information

3. Knowledge technologies and political participation

Threats: Political manipulation and surveillance
Opportunities: Enhanced democratic participation

4. Knowledge technologies and labour

Threats: Job displacement and economic inequality
Opportunities: Workers’ emancipation and labour rights

5. Knowledge technologies and human autonomy

Threats: Erosion of human autonomy / Lack of transparency and accountability
Opportunities: Removing human errors, biases to achieve fairer, more effective results

6. Knowledge technology and human identity

Threats: Bias and discrimination / Essentialism and conformism
Opportunities: self-expression and self-realisation

7. Knowledge technologies human creativity and expression

Threats: Standardisation + economic exploitation / Loss of control over knowledge creation and ownership
Opportunities: Machine-enhanced creativity / Democratisation of art

8. Knowledge technology and community building

Threats: Violence and discrimination/ Loss of trust
Opportunities: Networks of solidarity

9. KTs and the work of imagination

Threats: Misrepresentation and technophobia/ Perpetuating discrimination

Opportunities: Imagining alternative futures/ Anticipating issues

A visual map of the above issues and topics is available at this link: https://kings.padlet.org/eleonoralima/kt4d-conceptual-framework-tcd-s47m9gdlfd6coiw7.

3.3.4 Research questions

Once we identify our case studies, each section will address the following questions:

- Is the difference between past KTs and AI and big data a matter of substance or just scale?
- How did past examples of KTs shape and enhance democratic participation and human agency?
- What can be learned from these precedents? Are they still applicable after considering changes in our personal and societal values? To what extent?
- Did past examples of KTs lead to oppressive and antidemocratic systems and reduced human agency?
- What can be learned from these precedents? How did people respond and with what results?
- Historically, which groups of people (politicians, activists, artists, citizen associations, etc.) or institutions petitioned for a democratic use of KTs? Who were the groups historically left out from this progress/benefits?

3.3.5 Risks' identification and management

There are a number of risks that need to be considered for Module C, such as:

1. The model inferred from past knowledge technologies might not be applicable to the present, because of major technological and societal transformations that have occurred since;
2. Past knowledge technologies might not provide suitable models because the system of values in place at the time of their diffusion is now outdated and their biases and shortcomings are inherent to their specific historical context (both in terms of societal values, and of technological applications);
3. The project determines that the role of culture is more elusive than expected or that perhaps cultural norms and processes are too contingent to the historical, geographical, and social context and thus impossible to be subsumed under a general analysis.

If risks 1 and 2 materialise, we will then refocus the investigation so as to understand which fundamental aspects have changed over time (societal, political, cultural, technological), why they have changed and, finally, whether the change(s) constitute(s) progress or rather a tendency worth opposing. This will shed light on the trajectory of the relationship between knowledge technologies and people's sense of trust and free will, as well as on democratic participation more in general. Were risk 3 to materialise, it would nonetheless be a useful—if disappointing—conclusion. It would clarify the relationship between cultural norms and processes, and technological development, potentially supporting the view that, after all, cultural and technological forces operate on different levels – the first on the local one, the second on a global scale – so that aiming at capturing this entanglement in its totality amounts to erase the very cultural specificities that one aimed at represent.

3.4 Advancement of the state of the art

3.4.1 Initial case studies

In the table below, there are some initial examples of issues in AI and big data that we intend to analyse in context with past examples of KTs. This provides examples of the case studies considered and the list is not at all exhaustive.

General issue	**AI and big data**	**Historical precedent**
Psychological manipulation and personal freedom	Recommendation engines in dating apps and people's awareness (and concern) of being spied on and manipulated. People develop 'algorithmic awareness', and sense that the selection of suitable partners presented to them is guided by assumptions of who they are and what they (supposedly) like. Therefore, several people adopt measures in order to break the filter bubble of the recommendation engine so as to get 'unfiltered' results and preserve their freedom of choice and a sense of serendipity.	Concerns about the intrusiveness of the telephone in romantic conversations in the late 19th-early 20th century. E.g. couples knew that the phone operator could be listening in; chances for predatory men to have access to young women; fears that the connection was less real because not in person. This is what Carolyn Marvin (1990) defines as 'electric courtship': a new way of connecting romantically with people with the intrusion of a technological actor (the telephone).
Human agency: loss of control over knowledge creation and ownership	Generative AI tools are feared to replace human knowledge with algorithmically generated ones. For instance, today ChatGPT is trained on human-created data (e.g. coding samples on Stack Overflow), but the more people use it, the more the totality of our knowledge will be machine-created (is this a Ship of Theseus kind of situation?). How have people dealt with similar issues in the past, when new KTs seemed to take over tasks previously considered to be inherently human?	Over centuries, people have dealt with the recurring fear of losing the ability to memorise things, therefore of 'owning' knowledge. This was the first concern about the introduction of written language (Socrates),[21] and with the Internet (see: Umberto Eco). For example, for Primo Levi while in Auschwitz the ability to recite verses from Dante's *Inferno* that he memorised was the only moment he felt human, because he 'possessed' poetry and beauty inside himself (see *If this is a man*).
Human agency: loss of control over knowledge creation and ownership	There are concerns around the use of AI to generate content to support political and social causes. For instance, fashion companies are using AI generated models to have a more diverse cohort of people promoting their clothes. A so-called 'AI influencer' shared a story about a case of sexual harassment, that logically never happened, to spread awareness. On May 1, 2023, Amnesty published on Twitter a (new deleted) AI-generated picture representing police brutality in Colombia. What does it mean to delegate political activism to AI? Is it a way to distance ourselves from 'real' people and violence?	How is the case of Amnesty's AI generated picture different from picture manipulation? What are cases in which blurring the line between style and content upset people for the political statement underneath? In 2011 a war photographer used the iPhone filter Hipstamatic to shoot US soldiers in Afghanistan. People were concerned about how using this filter made the pictures look older, more distant, and created a sort of 'simulated nostalgia'. Also, the final pictures were not the product of the photographer's gaze, but a random result decided by the app.
Psychological manipulation: human and textual agency	Chatbots, especially therapy and company ones, pose some crucial questions around people's agency and the risk of psychological manipulation (e.g. people feeling romantically involved with bots, the	What kind of agency have people attributed to texts (written, audio, visuals) over the centuries? With what implications for human agency? For instance, Greek tragedy and romantic novels used to be often accused of inciting overwhelming emotions and thus depriving people of their rational thinking (e.g. people reading

[21] John Hollander writes: "The notorious charge levelled by Socrates in the Phaedrus against the technology of writing, and how inventing it supplanted and ruined the earlier, better, and somehow more natural operations of memory [...] suggests that the very invention of writing was a new technology whose product would be what we call literature" (306).

	case of the person who committed suicide after having interacted with a therapy chatbot).	Goethe's *Werther* and then committing suicide to emulate the protagonist).
KTs and political participation	The examples of how AI and big data are negatively impacted democratic participation are overwhelming. The impact that AI and Machine Learning can have on orienting people's opinion has become evident, for instance, during the 2016 US Presidential election (Guglielmi) and the UK Brexit referendum (Bastos and Mercea), when Russian operatives used bots—AI automated accounts that share content—to spread fake news on social media in order to influence the electorate. Furthermore, AI employed by governments to control their citizens, such as in the cases of the profiling of the Uighur population by the Chinese government (Mozur) or during the 2019-2020 protests in Hong Kong (Fussell).	While it would be very easy to identify a precedent of the same negative impact of KTs over democratic participation, it would be more interesting and useful to focus on case studies that point in a different direction. When did KTs enhance democratic participation and how? Under what circumstances? For instance, in Czechoslovakia after the Great Purge was denounced by Soviet leader Nikita Khrushchev following Stalin's death, television became a powerful tool for democratic participation and played a central role in the Prague Spring (Bren). A similar example is the role of the website for citizen journalism Indymedia in the late 90s-early 2000s.
Human creativity and expression	AI systems and technologies are increasingly being used to generate artworks that are indistinguishable from those made by human artists. AI-generated artworks raise questions of artistic creativity and human consciousness. E.g., on 17 May 2023, Sudowrite launched Story Engine, an AI tool for writing long-form stories. This sparked a heated debate among writers and readers: why are we building systems doing the things we enjoy? Shouldn't creativity be entirely *for us by us*?	It is possible to contextualise the current involvement of AI in creative endeavour by looking not only at the long history of computer-generated (the first poem written by a computer dates back to 1959), but at earlier discussion on the threats and opportunities of relinquishing human control over authorship. Ràmon Llull's in his treatise *Ars Magna* (1305), set the theory for a paper machine, known as the 'Llullian Circle' meant to use the combinatorial process to produce text. In 1937, Borges wrote a Text called "Ramon Llull' s Thinking Machine."
Job insecurity and labour rights	The impact of AI and big data on workers' rights is the focus of ongoing concerns. People fear to be made redundant due to the automatisation of many tasks (which more and more include intellectual jobs as well). The issues of workers' exploitation and invisibility is also central when thinking of crowdsourcing marketplaces like Amazon Mechanical Turk (Irani and Silberman), or in the case of content moderators and workers labelling data who are not only underpaid, but also deal with emotional trauma due to the content view.	The first 'strike against automation' took place in Coventry UK between April and May 1956, when Standard Motor Company workers began an industrial dispute aimed at preventing the dismissal of 3,000 workers in consequence of the introduction of automated methods of production (Castoriadis: 26-27). Around those years, many debates (UK and US government, Soviet Union, workers unions) and publications discussed the automation of work and its implications for workers' rights.
Possibilities for self-expression	AI and big data can allow for greater customisation, which could in turn lead to better representation and greater freedom for their users. While this has been used mostly for economic gains (e.g. targeted advertisement), there is space to use these technologies for good. One example is the use of AI technologies to allow for greater access for people with disabilities (Wald). Blind or visually impaired people could be able to "drive" autonomous cars, and artists with motor disabilities are already taking advantage of some AI-powered tools.	Historically, new technological tools have often led to a more truthful, effective expression of one's identity and thoughts. This is something that all KTs have promised to people over the centuries, as they allowed for a more 'immediate' (Bolter and Grusin), direct and genuine form of expression. This was true for writing (personal take on stories rather than epic poems that are memorised and recited), for photography, for independent radios and TV channels, for the Internet. What positive examples can we identify, and can we learn from them?

3.4.2 Example of comparing past and present KTs: ChatGPT and the printing press

This is an example of how Module C will address specific instances and issues instead of considering past KTs (in this case the printing press) as a whole.

Currently there is a pressing need to verify our sources of information, not just to distinguish between reliable and unreliable sources, but between human-created and synthetic data. This has become even more pressing since the release of the new version of ChatGPT. Furthermore, algorithms are not only in charge of producing but also of certifying knowledge. In this respect, Gillespie (2016) writes: “That we are now turning to algorithms to identify what we need to know is as momentous as having relied on credentialed experts, the scientific method, common sense, or the word of God.”

Can we identify moments in history when similar changes in who has authority over producing and certifying knowledge occurred? What were the political ramifications back then, and what we can learn from it?

For instance, in the 16thCentury, because of the exponential growth of information available due to the advent of the printing press, people needed to develop new tools and practices to discriminate between their sources of knowledge, as the old structure (e.g. monasteries and universities that produced manuscripts and guaranteed of their quality) quickly disappeared (Blair 2010). Like today, there was a rapid and unprecedented increase in the data being produced, and no structure in place to verify their reliability. Two were the main solutions: one from the bottom down, one from the bottom up.

First, only political authorities such as the emperor, the local Government, the Pope could grant the licence to print, which led to an imbalance in people’s access to knowledge: books printed in the Republic of Venice or in the Netherlands, which were relatively free-thinking places, were more reliable as they did not undergo censorship like books printed, for example, in the Vatican State (Grendler; Sachet).

However, printers also took upon themselves to develop a way to reassure their customers about the quality of their product (and the reliability of the sources). Therefore, each printer developed a printer's mark which functioned as a trademark (Wolkenhauer and Scholz). These became extremely important as they provided information about who and where a book was printed (a reputable printer? A free-thinking country?) and indeed to this day scholars who work on early printed texts need to be knowledgeable of this system. This, as expected, led to many cases of forgery, as less reputable printers counterfeited printer’s mark from more respected workshops, located in (relatively) censor-free countries. This nullified the governments’ attempt to certify the good quality of the sources printed in their own countries.

To what extent does this situation resemble the proposal of the UK government to build a British version of ChatGPT (Hern) to exert more control over LLMs and provide a guaranteed certificate to their citizens? What are the consequences of addressing these issues by creating enclaves in which rules and regulations differ? Who will not benefit from it?

How can the printers’ marks and their commercial values help us understand better something like the discussion on Twitter’s blue checkmark, which can now be simply purchased by anyone?

3.5 References

No	Description/Link
R1	Aamodt, Agnar and Mads Nygard. “Different Roles and Mutual Dependencies of Data, Information, and Knowledge – An AI perspective of Their Integration.” *Data & Knowledge Engineering*, Vol. 16, 1995, pp. 191-222.
R2	Adams, Rachel. “Can Artificial Intelligence Be Decolonized?” *Interdisciplinary Science Reviews*, Vol. 46 No. 1-2, 2021, pp. 176-197.
R3	Adamson, Greg and R. Kline, K. Michael, M. G. Michael. “Wiener’s Cybernetics Legacy and the Growing Need for the Interdisciplinary Approach.” *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)*, Vol. 103, No. 11, November 2015, pp. 2208-2214.
R4	Alarie, Benjamin and A. Niblett, A. H. Yoon. “How Artificial Intelligence Will Affect the Practice of Law.” *University of Toronto Law Journal*, Vol. 68, No. 1, January 2018, pp. 106-124.
R5	Ananny, Mike. “Seeing Like an Algorithmic Error: What are Algorithmic Mistakes, Why Do They Matter, How Might They Be Public Problems?” *Yale Journal of Law and Technology White Paper Series*. Vol. 14, 2022, pp. 1-21.
R6	Anderson, Jackson T. and J. Freybote, D. Lucus, M. J. Seiler, L. Simon. “Using Artificial Intelligence to Identify Strategic Mortgage Default Attitudes.” *Journal of Real Estate Research*, 22 December 2021.
R7	Asimov, Isaac. *I, Robot*. Gnome Press, 1950.
R8	Bastos, Marco T., D. Mercea. “The Brexit Botnet and User-Generated Hyperpartisan News.” *Social Science Computer Review*, Vol. 37, No. 1, February 2019, pp. 38–54.
R9	Bell, Genevieve and M. Gould, B. Martin, A. Mclennan, E. O’Brien. “Do More Data Equal More Truth? Toward a Cybernetic Approach to Data.” *The Australian Journal of Social Issues*, Vol. 56, No. 2, May 2021.
R10	Blackwell, Alan F. and Addisu Damena, Tesfa Tegegne. “Inventing Artificial Intelligence in Ethiopia.” *Interdisciplinary Science Reviews*, Vol. 46 No. 3, 2021, 363-385. DOI: 10.1080/03080188.2020.1830234.
R11	Blair, Ann. “Reading Strategies for Coping with Information Overload ca. 1550-1700. *Journal of the History of Ideas*, Vol. 64, No. 1, January 2003, pp. 11-28.
R12	Blair, Ann. *Too Much to Know: Managing Scholarly Information Before the Modern Age*. Yale University Press, 2010.
R13	Boellstorff, Tom. *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. Princeton University Press, 2015.
R14	Bolter, Jay David and Richard Gruisin. *Remediation: Understanding New Media*. MIT Press, 1999.
R15	Borghoff, Uwe M. and Remo Pareschi (eds.), *Information Technology for Knowledge Management*. Springer, 1998. ISBN: 9783642083563.
R16	Bren, Paulina. *The Greengrocer and His TV: The Culture of Communism After the 1968 Prague Spring.* Cornell University Press, 2010.
R17	Callon, Michel and Arie Rip, John Law, *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. Palgrave, 1986.
R18	Castoriadis, Cornelius. *Political and Social Writings: Volume 2, 1955–1960. From the Workers’ Struggle Against Bureaucracy to Revolution in the Age of Modern Capitalism*. University of Minnesota Press, 1988.

Funded by the European Union

KT4D has received funding from the EU’s Horizon Europe research and innovation programme under Grant Agreement no. 101094302.

R19	Cavgias, Alexsandros and Raphael Corbi, Luis Meloni, Lucas M. Novaes. “Media Manipulation in Young Democracies: Evidence From the 1989 Brazilian Presidential Election.” *Comparative Political Studies*, 2023.
R20	Chávez Heras, Daniel and Tobias Blanke. “On Machine Vision and Photographic Imagination.” *AI & Society*, No. 36, 2020, pp. 1153–1165.
R21	Crenshaw, Kimberlé. “Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Anti-discrimination Doctrine Feminist Theory and Antiracist Politics.” *University of Chicago Legal Forum*, No. 9, 1989, pp. 139-167.
R22	de Swarte, Thibault and Omar Boufous, Paul Escalle. “Artificial Intelligence, Ethics and Human Values: The Cases of Military Drones and Companion Robots.” *Artificial Life and Robotics*, vol. 24, no. 3, 2019, pp. 291–296.
R23	Dionisio-Flores, Hernan and Méndez Vergaray, Juan, Jaysson, Dennys, Durand, Picho, Farfán Pimentel, Johnny Félix, Masías, Edward. “Learning and Knowledge Technologies in School Management in Times of Covid-19.” *Theoretical Review. Drugs and Cell Therapies in Hematology*. Vol. 10 No. 1, July 2021, pp. 522-527.
R24	Ellul, Jacques. *The Technological Society* (1954). Random House, 1964.
R25	Franklin, Ursula. *The Real World of Technology* (1990). House of Anansi Press, 2004.
R26	Fussell, Sidney. “Why Hong Kongers Are Toppling Lampposts.” *The Atlantic*, 30 August 2019.
R27	Galison, Peter. “The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision.” *Critical Inquiry*, Vol. 21, No. 1, Autumn, 1994, pp. 228-266.
R28	Garavelli, Claudio and Michele Gorgoglione, Barbara Scozzi. “Managing Knowledge Transfer by Knowledge Technologies.” *Technovation*, Vol. 22, No. 5, 2002, pp. 269-279.
R29	Gillespie, Tarleton. “Algorithm.” In Peters, Ben (ed.). *Digital Keywords: A Vocabulary of Information Society and Culture*. Princeton University Press, 2016.
R30	Gitelman, Lisa. *Always Already New: Media, History, and the Data of Culture*. MIT Press, 2006.
R31	Greenaway, Naomi. “The Ethics of Spying on Your Children.” *The Telegraph*. 10 August 2023.
R32	Grendler, Paul F. “The Roman Inquisition and the Venetian Press, 1540-1605.” *The Journal of Modern History*, vol. 47, no. 1, 1975, pp. 48–65.
R33	Guarino, Nicola. “Formal Ontology, Conceptual Analysis and Knowledge Representation.” *International Journal of Human-Computer Studies*, Vol. 43, 1995, pp. 625-640.
R34	Guglielmi, Giorgia. “The Next-Generation Bots Interfering with the US Election.” *Nature,* vol. 587, no. 21, 20 October 2020.
R35	Habgood-Coote, Joshua. “Deepfakes and the Epistemic Apocalypse.” *Synthese*, Vol. 201, No. 3, 2023, pp. 1-23.
R36	Hao, Karen. “The Problems AI Has Today Go Back Centuries.” *MIT Technology Review*, July 31, 2020.
R37	Hern, Alex. “UK Needs its Own ‘BritGPT’ or Will Face an Uncertain Future, MPs Hear.” *The Guardian*, 22 February 2023.
R38	Hofstede, Geert. “Dimensionalizing Cultures: The Hofstede Model in Context.” *Online Readings in Psychology and Culture*, Vol. 2 No. 1, 2011.
R39	Hollander, John. “Literature and Technology: Nature’s ‘Lawful Offspring in Man’s Art’,” in Mack, Arien (ed.). *Technology and the Rest of Culture*. Ohio State University Press, 1997, pp. 305-330.
R40	Holtshouse, Dan K. “Foreword.” In: Borghoff, Uwe M. and Remo Pareschi (eds.). *Information Technology for Knowledge Management*. Springer, 1998, pp. V-VI.

R41	Irani, Lilly C. and M. Six Silberman. "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, 2013, pp. 611–620.
R42	Johnston, John. *The Allure of Machinic Life: Cybernetics, Artificial Life, and the New AI*. MIT Press, 2008.
R43	Kelly, Heather. "Parents Are Using AirTags to Track Kids and Give Them Freedom." *The Washington Post*. 27 July 2023.
R44	Kember, Sarah and Joanna Zylinska. *Life after New Media: Mediation as a Vital Process*. MIT Press, 2012.
R45	Khan, Shahnawaz and Mustafa Raza Rabbani. "In-Depth Analysis of Blockchain, Cryptocurrency and Sharia Compliance". *International Business Innovation and Research,* Vol. 29. No. 1, 2022, pp. 1-15.
R46	Kittler, Friedrich. "There Is No Software." In Johnston, John (ed.). *Literature, Media, Information Systems*. Overseas Publishers Association, 1997, pp. 147–155.
R47	Koggalahewa, Darshika and J. L. Amararachchi, S. U. Pilapitiya and D. T. K. Geegange. "Semantic Self Learning and Teaching Agent (SESLATA)." *2013 8th International Conference on Computer Science & Education*, Colombo, Sri Lanka, 2013, pp. 171-176.
R48	Kosinski, Michal. "Facial Recognition Technology Can Expose Political Orientation from Naturalistic Facial Images." *Scientific Reports*, Vol. 11, No. 100, 25 November 2021.
R49	Lachner, Florian, and Constantin von Saucken, Florian Mueller, Udo Lindemann. "Cross-Cultural User Experience Design Helping Product Designers to Consider Cultural Differences." In: Rau, P. (ed.). *Cross-Cultural Design Methods, Practice and Impact*. *CCD 2015. Lecture Notes in Computer Science*, Vol. 9180, 2015.
R50	Laffier, Jennifer and Rehman, Aalyia. "Deepfakes and Harm to Women." *Journal of Digital Life and Learning*. Vol. 3 No. 1, 2023, pp. 1-21.
R50	Latour, Bruno. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, 2005.
R51	Leslie, David et al. "Does 'AI' Stand for Augmenting Inequality in the Era of Covid-19 Healthcare?" *BMJ (Clinical research ed.),* Vol. 372 No. 304, 15 March 2021.
R52	Levesque, Hector J. "Knowledge Representation and Reasoning." *Annual Review of Computer Science*, Vol. 1, 1986, pp. 255-287.
R53	Mackenzie, Donald and Judy Wajcman, *The Social Shaping of Technology: How the Refrigerator Got its Hum*. Open University Press, 1985.
R54	Manovich, Lev. "New Media from Borges to HTML." In: Wardrip-Fruin, Noah and Nick Montfort (eds.) *The New Media Reader*. MIT Press, 2003, pp. 13-25.
R55	Manovich, Lev. *The Language of New Media*. MIT Press, 2002.
R56	Marvin, Carolyn. *When Old Technologies Were New: Thinking about Electric Communication in the Late Nineteenth Century*. Oxford University Press, 1990.
R57	McLuhan, Marshall. *The Gutenberg Galaxy: The Making of Typographic Man* (1962). University of Toronto Press, 2011.
R58	McLuhan, Marshall. *Understanding Media: The Extension of Man* (1964). MIT Press, 1994.
R59	McSweeney, Brendan. "The Essentials of Scholarship: A Reply to Geert Hofstede." *Human Relations*, Vol. *55* No. 11, 2002, pp. 1363-1372.
R60	Milton, Nicholas Ross. *Knowledge Technologies*. Polimetrica, 2008.
R61	Moussetes, Agneta. "The Absence of Women's Voices in Hofstede's Cultural Consequences: A Postcolonial Reading." *Women in Management Review*, No. 22, 2007, pp. 443–445.

R62	Mozur, Paul. “One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority.” *The New York Times*, 14 April 2019.
R63	Pangaro, Paul. “Cybernetics as Phoenix: Why Ashes, What New Life?”. In Werner, Liss C. (ed.). *Cybernetics: State of the Art*. Universitätsverlag TU Ber, 2017.
R64	Pickering, Andrew. “Ontology and Antidisciplinarity.” In: Barry, Andrew and Georgina Born (eds.). *Interdisciplinarity: Reconfigurations of the Social and Natural Sciences*, Routledge, 2013, pp. 209-225.
R65	Pickering, Andrew. *The Adaptive Brain: Sketches from Another Future*. Chicago University Press, 2010.
R66	Plato, *Phaedrus*. In: Plato. *Euthyphro. Apology. Crito. Phaedo. Phaedrus*, edited by Harold North Fowler. Harvard University Press, 1999, pp. 405-580.
R67	Postman, Neil. *Technopoly: The Surrender of Culture to Technology*. Knopf, 1992.
R68	Ragnedda, Massimo and Glenn Muschert. “The Political Use of Fear and News Reporting in Italy: The Case of Berlusconi’s Media Control.” *Journal of Communications Research*, Vol. 2, No. 1, 2010, pp. 43-54.
R69	Rheingold, Howard. *Tools for Thought: The History and Future of Mind-Expanding Technology* (1985). MIT Press, 2000.
R70	Rigau, German and Bernardo Magnini, Eneko Agirre, P. Vossen, John A. Carroll. “MEANING: a Roadmap to Knowledge Technologies.” *RAODMAP@COLING*, 2002.
R71	Sachet, Paolo. *Publishing for the Popes: The Roman Curia and the Use of Printing (1527–1555)*. Brill, 2020.
R72	Schwab, Klaus. *The Fourth Industrial Revolution*. Penguin Random House, 2016.
R73	Stein, Anthony and Sven Tomforde, Ada Diaconescu, Jörg Hähner, Christin Müller-Schloer. “A Concept for Proactive Knowledge Construction in Self-Learning Autonomous Systems." *2018 IEEE 3rd International Workshops on Foundations and Applications of Self Systems*, Trento, Italy, 2018, pp. 204-213. doi: 10.1109/FAS-W.2018.00048.
R74	Stewart, Cherry and Khan, Ashfaq Ahmad. “A Strategy for Using Digital Mindsets and Knowledge Technologies to Move Past Pandemic Conditions.” *Accounting Research Journal*, Vol. 34 No. 3, 2021, pp. 345-356.
R75	Vidales, Carlos. “Cybernetics and Its Conceptual Legacy.” *Cybernetics and Human Knowing*, Vol. 27, No. 3, 2020, pp. 5-8.
R76	Wald, Mike. “AI Data-Driven Personalisation and Disability Inclusion.” *Frontiers in Artificial Intelligence*, Vol. 3, 2020.
R77	West, Emily. “Amazon: Surveillance as a Service.” *Surveillance & Society*, Vol.17, No. 1/2, 2019, pp. 27-33.
R78	Wiener, Norbert. *The Human Use of Human Beings: Cybernetics and Society* (1950). Sphere Books, 1968.
R79	Williams, Raymond. *Television: Technology and Cultural Form*, Fontana, 1971.
R80	Wolkenhauer, Anja and Bernhard F. Scholz (eds.). *Typographorum Emblemata: The Printer’s Mark in the Context of Early Modern Culture*. De Gruyter, 2018.
R81	Wood, Alex J., and M. Graham, V. Lehdonvirta, I. Hjorth. “Good Gig, Bad Gig: Autonomy and Algorithmic Control in the Global Gig Economy.” *Work, Employment and Society*, Vol. 33, No. 1, February 2019, pp. 56-75.
R82	Zhuang, Yue-ting and Fei Wu, Chun Chen, Yun-he Pan. “Challenges and Opportunities: From Big Data to Knowledge in AI 2.0.” *Frontiers of Information Technology & Electronic Engineering*, Vol. 18 No. 1, 2017, pp. 3-14.