Can an AI Analyze Arguments? Argument-Checking and the Challenges of Assessing the Quality of Online Information

Brave , R.; Russo, F.; Uzovic, O.; Wagemans, J.

DOI
10.1201/9781003261247-20

Publication date
2023

Document Version
Final published version

Published in
AI and Society

Published in:
Christo El Morr (eds)
AI and Society. Tensions and Opportunities
CRC Press, Taylor & Francis Group

CHAPTER 16

Can an AI Analyze Arguments? Argument-Checking and the Challenges of Assessing the Quality of Online Information

Ruben Brave[a], Federica Russo[b], Ondrej Uzovic[c], and Jean Wagemans[b]

[a]*CEO Entelligence.nl and Co-founder Internet Society NL MMGA Working Group, Amsterdam, The Netherlands*
[b]*Faculty of Humanities, University of Amsterdam, Amsterdam, The Netherlands*
[c]*Independent Software Engineer, Bratislava, Slovakia*

CONTENTS

INFORMATION OVERLOAD IN THE DIGITAL ERA

Information and communication technologies (ICTs) profoundly transformed science and everyday life. From the use of digital computers to big data, from re-shaping social interactions and significantly altering the formation and perception of the self

DOI: 10.1201/9781003261247-20

(Zandbergen, 2011), it is not an overstatement to say that ICTs mark a revolution, the *digital* revolution. Specifically, technological developments in information and communication distribution and processing have increased the speed and amount of information shared. This has happened already in the past, in the shift from pre-history to history, and not to hyper-history. The information cycle of occurrence, transmission, process, management, and use of information underwent significant changes through time: in pre-history, we were not able to record information, with oral transmission of knowledge. The advent of writing made us enter a new phase, that of history. We live now in the *zettabyte* era. Digital technologies have changed the landscape, marking the beginning of hyper-history. However, the difference between history and hyper-history lies not merely in the quantity and speed of information that is transmitted, but mostly in *how* it is transmitted and processed. The hallmark of ICTs since the digital revolution is creating, using, and rediscovering *connections* (Floridi, 2014, 2015). Digital technologies allow for many more *connections* to be made, and it is in this sense the concepts such as "speed of evolution of knowledge" and "collective intelligence" (Lévy, 1997) or as "connective intelligence" (De Kerckhove, 1998) have been introduced.

It is often assumed that the *amount* of information circulating is the root cause of other problems, and notably of the *quality* of information circulating online. Some authors have challenged this premise (Altay et al., 2021). In this chapter, we don't directly engage with this dispute about quantity vs quality of information. We take it that the sheer amount of information human internet users can handle poses important challenges for our material capacities in selecting and assessing it, also for time constraints. However, in this chapter, we focus on aspects related to the *quality of online information*. In particular, we focus on the handling of information internet users are confronted with, as well as currently available solutions or coping strategies (Wardle & Derakhshan, 2017). There is a wealth of research done and in progress about the notion of information overload and on how individuals react to it. Some studies model interactions among individuals in online spaces (Jones et al., 2004; White & Dorman, 2000); other studies focus on how information overload affects our experience as consumers (Li, 2017) or as health information seekers (Swar et al., 2017); yet others look at our social interactions and interpersonal trust (Beaudoin, 2008; Ellwart et al., 2015) or at the influence of information overload on how we see ourselves (Palfrey & Gasser, 2016) and our relationship to knowledge (Wardle & Derakhshan, 2017) and news consumption (Marwick & Lewis, 2017).

Ultimately the problem of online information quality is that it is difficult to establish what is true, and who or what is a reliable source, and this is the case whether we are overloaded with "good" or with "bad" information. While these are well-known and studied problems (Borg, 2019; Brave, n.d.; MMGA, 2019; Roetzel, 2019), in this paper we take a look at the problem of online information quality from the perspective of argumentation theory and artificial intelligence. In particular, we present and discuss an ongoing project to develop a glass-box AI engine called KRINO – from Greek, to judge, criticize, reason – capable of parsing written text on the discourse level and analyzing the arguments thereby contained. KRINO is designed to assist users with argument-checking, i.e., the process of

analyzing the characteristics of arguments in order to be able to apply domain and user-specific criteria for assessing them. We describe the set-up and basic features of KRINO and explain how it can assist human annotators in a project undertaken by the Dutch organization Internet Society Netherlands Make Media Great Again Working Group (*shortened* MMGA) that is aimed at improving the quality of online information in settings varying from online news outlets to social media. The joint project is motivated by the need to empower internet users to better analyze online information and to distinguish between "good" and "bad" information, or between information, mis- and dis- and mal-information. We explain the prospects and challenges of combining the KRINO and MMGA projects on argument-checking and discuss the societal and computational relevance of this project.

The chapter is structured as follows. In Section 2, we discuss fact-checking, a valuable activity often presented as the latest frontier to fight mis-, dis-, and mal-information. We explain why, while valuable, fact-checking is not enough to address the information overload and to improve upon the quality of online information. In Section 3, we introduce argument-checking as a distinct approach to argumentation, flexible and agile, able to be used in a variety of settings and by individuals with varying levels of education and expertise to check the quality of pieces of online information. In Section 4, we present work in progress to support the process of argument-checking with a glass-box AI engine called KRINO. The project of developing an AI able to analyze arguments is motivated by our specific take on the problem of poor quality of online information, and the prospects of argument-checking to address it. In Section 5, we conclude the chapter with a reflection on the societal relevance of argument-checking and of KRINO.

FROM FACT-CHECKING TO ARGUMENT-CHECKING

Fact-checking isn't a new phenomenon, and can be rightly considered as a key journalistic action since at least the 1920s (Fabry, 2017); nowadays, we consider its mission to debunk false statements, especially in politics, but not only. There are several kinds of organizations involved in fact-checking, and world-wide. Fact-checking involves numerous professional figures, also outside journalism, and it is growing in proposing approaches and methods to select and then assess claims made in the public sphere.

As a process, fact-checking seeks to verify presented information (e.g. text, video, sound) in order to promote conformity to facts and correctness of reporting. Fact-checking can be conducted before (*ex ante*) or after (*ex post*) the information is published or otherwise disseminated. While *ex ante* fact-checking aims to identify errors so that the information can be corrected or even rejected before dissemination, *ex post* fact-checking is often followed by a written or visual report of inaccuracies. Internal fact-checking is part of the regular journalistic process and is done in-house by the publisher; in case the presented information is analyzed by a third party, the process is categorized as external fact-checking (Graves & Amazeen, 2019). Examples of organizations devoted to the latter are FactCheck.org and PolitiFact in the US and Full Fact in the UK. This type of fact-checking first emerged in the US in the early 2000s and, after it grew in relevance, started to spread to other countries (Graves & Amazeen, 2019).

Fact-checking and its methodologies are increasingly the subject of academic and non-academic evaluation. While generally considered a valuable activity, it has also been criticized, not only for employing questionable methodology regarding the selection of statements and the choice of criteria for evaluating them but also for its limited effectiveness in fighting mis- and disinformation (see, for instance, (Barrera Rodriguez et al., 2017; Nyhan & Reifler, 2010, 2012; Thorson, 2016; Uscinski & Butler, 2013; Wintersieck, 2017)).

To some extent, fact-checking indeed seems to correct perceptions among citizens (Drutman, 2020) although it depends on the way it is conducted (Clayton et al., 2020). The importance of fact-checking notwithstanding, its effectiveness is also under close scrutiny, with important considerations about whether the effects of debunking information last long enough, which groups are more or less susceptible to change their beliefs or not, etc. (Nyhan 2021; Porter and Wood 2021). As information overload is real, the risk of repeatedly being exposed to fake news is only to be expected. This can increase the perceived truthfulness of fake news (Pennycook et al., 2018). Actually, the mere fact that people encounter a specific fact-check frequently can create distorted memories of the veracity of false claims, the "illusion of truth" effect (Skurnik et al., 2005). Overall, the current status seems to be that the correctional impact of fact-checking on people's beliefs is questionable because the effectiveness is influenced by preexisting beliefs, ideology, and knowledge on the side of the information receiver (Walter et al., 2020).

Apart from these criticisms, it has been observed there is a significant limitation of the scope of fact-checking in that it only evaluates the truth of isolated statements of fact. As Plug and Wagemans (2020, pp. 236–237) put it:

> Independent of their being true or false, statements of fact may fulfill an argumentative function in the discourse, in which case they are put forward to establish or increase the acceptability of the arguer's point of view. […] [The] scope [of fact-checking] is relatively limited in that it only involves the assessment of the truth of an isolated statement of fact. It does not address the argumentative relationship between that statement and the claim it intends to support, nor any other aspects of the rhetorical design of the discourse.

Although we acknowledge the relevance and merits of fact-checking (and of the scholarship that studies it), we think it can be supplemented with argument-checking to drive a substantial change in improving the quality of online information. As anticipated above, fact-checking doesn't cover all aspects of debunking misinformation and fake news. Statements of fact are often used to support the acceptability of other claims, which can be statements of fact, value or policy – see Plug and Wagemans (2020, pp. 245–49) for analyses of examples of the ways in which statements of fact can be embedded in arguments. Given this embeddedness, many problems regarding the quality of information remain outside of the limited scope of fact-checking: empirical statements expressing correct (or roughly correct) facts can be used in bad reasoning. All in all, in a good argument, there is more than correctness of the facts. For these reasons, verifying the quality of arguments themselves

seems a necessary and fundamental part of the information quality control process, which is the approach we present in the next section.

ARGUMENT-CHECKING AND ONLINE INFORMATION

In this section, we present our approach to argument-checking, qua human annotation. Different from most formal logical approaches, our approach is suitable for individuals of varying educational levels and enables people to analyze and evaluate *natural* arguments, i.e., arguments expressed in natural language and encountered in their everyday lives, for instance on social media, websites, or any type of online platform.

The activity of argument-checking requires a set of skills or competencies for interpreting persuasive discourse, whether that is a single persuasive message or a complete text aimed at convincing the reader to believe something or to do something (Wagemans, Forthcoming). Among these competences are, first of all, "argument detection", i.e., finding out what the main claim is that the author of the discourse wants to convey to their audience and which arguments have been put forward in support of that claim. Then, the reader or listener must find out how the argumentative elements contained in the discourse hang together, thus creating a structured picture of its argumentative fabric. This competence can be called "argument mapping". Further, in order to be able to judge the quality of the argumentation, one would need to zoom in on the individual arguments on the map and study the relationships of support between the main claim and the chains of argument put forward in support of it. Guidelines for this activity of "argument type identification" have been developed in the so-called Argument Type Identification Procedure (ATIP) (Wagemans, 2021). Once it has become clear what types of arguments are represented in the text or discussion, "argument assessment" can take place by asking specific critical questions relevant to their evaluation. To assist the analyst in this final task, specific evaluation procedures have been developed such as the Comprehensive Assessment Procedure for Natural Argument (CAPNA) (Hinton & Wagemans, 2022).

The activity of argument-checking can thus be divided into a chain of smaller activities, with the output of the previous link in that chain functioning as the input of the next: subsequently, the arguments are detected, they are mapped, their type is identified, and they are evaluated. Each of the individual links in this sequence requires different competences. The level that can be reached in acquiring these competences may vary among individuals, relying on an interdependent cluster of factors: their ability to recognize reasoning expressed in language, their knowledge of rhetorical strategies for producing argumentative discourse, and the length and intensity of relevant experience in processing, understanding, and assessing the quality of such discourse. Our procedural approach to argument evaluation, however, enables the development of a fine-grained training program aimed at enhancing people's competences in specific (sub)skills of argument-checking. Training in argument-checking can happen at various levels of education (for instance, students at various stages of education, early careers in research, etc.), and be tailored to specific domains of application (for instance, argumentation in legal settings, or compliance in the automotive sector, evidence assessment in the health domain, etc.). Moreover, the above procedures can be automated to a certain extent and

implemented into the design of argument technology. This is because ATIP and CAPNA, unlike other approaches in argument evaluation, are quasi-algorithmic procedures by design and do not work with predetermined forms of valid arguments (Wagemans, 2020; Hinton & Wagemans, 2022).

This brings us to the basic idea behind our joint project, which is that, indirectly and in the long term, we can improve the quality of online information by increasing the literacy of individuals. By providing them with training in argument-checking and tools to help them perform such checking, we aim to "immunize" people to low levels of information quality and enable them to (pro)actively contribute to a better online information exchange. More specifically, the project is aimed at:

i. Increasing the literacy of individuals (as online *users*) to make themselves immune against the negative effects of dis- and mis-information.

ii. Empowering individuals (as online *agents*) to intervene and block in appropriate ways episodes of dis- and misinformation, of trolling, or other.

iii. Teaching individuals (as online content *producers*) to share and disseminate high quality information online.

iv. Certifying the (increased) level of critical thinking via a Comprehensive Measure of Argumentation Skills (CMAS).

We aim to develop a CMAS precisely to be able to continuously tailor and fine-tune training on argument-checking to specific target groups, with varying degrees of educational levels and with different domains of expertise and background knowledge. The course "From fact-checking to argument checking" part of the Honors Programme run at the Institute for Interdisciplinary Studies at the University of Amsterdam, and offered for three academic years beginning in 2021–22, is a first concrete step in this direction.

For accomplishing these aims, we take inspiration and guidance from the field of critical pedagogy (Freire et al., 2014; Knight et al., 2020). Critical pedagogy promotes a specific approach to education, and notably one in which we strive to *empower* students, citizens, and, in our case, users and producers of online content. Applied to argument-checking, the idea is to empower users and producers of online information by awakening their critical consciousness, and also by providing them with tools that they can put to use: argument-checking as a critical pedagogy approach to digital literacy (Brave et al., 2022).

Developing a theoretical, practical, and pedagogical approach to argument-checking is also part of a collaboration with MMGA, within which we are designing training programmes on argument-checking, tailored to different audiences. MMGA is a blockchain-based annotation platform (with hundreds of registrants) in which screened and trained expert and/or critical thinking readers can annotate high-impact news sites such as NU.nl and AD.nl, two of the "Big Four" largest Dutch online news platforms. MMGA has set up a collaboration between publishers and a screened community of readers, viewers, and listeners to jointly counteract the effects of misinformation and improve the quality of

media. To achieve this goal, MMGA has built a transparent system for actionable suggestions from this community pool, which functions as an annotation platform and has been tested on NU.nl, a major Dutch news outlet with 7–8 million visitors. The test involved a group of critical and knowledgeable NU.nl readers (called "annotators") who were motivated to critically assess journalism news articles; annotators received instructions and were checked for their capabilities before being allowed to annotate. They then offered suggestions to increase the journalistic quality through the balanced use of sources and clearer transfer of information (Brave, 2019, 2021).

The automatic detection of fake news through natural language processing, machine learning, and network analysis is high on the agenda of several tech enterprises (Islam et al., 2021). The main proposition is that autonomously working systems will be able to categorize information as "fake news" and help to decrease the probability of users encountering it (Pennycook & Rand, 2021). As we remarked above, the procedures for argument-checking can be (partially) automated and implemented into tools that can assist the human user in analyzing and evaluating the quality of online information. This is currently done in the KRINO project, which we present in the next section. In our view, the collaboration between MMGA, with its involvement of human annotators, and the developers of the KRINO AI engine strengthens the shared mission of reversing the trend of an increased amount of disinformation, fake news, and poor journalism that is progressively dividing the world and having more impactful societal and psychological consequences each day.

AUTOMATING ARGUMENT-CHECKING: THE KRINO PROJECT

The general aims of the collaboration between MMGA and KRINO are to develop argument-checking as a complementary activity to fact-checking, to have annotators rather than experts carry out this activity, and to help them do so by partially automating the process of analyzing and evaluating arguments (Nieman, 2020).

As we explained in Section 3, the activity of argument-checking requires various competences, some of which are more easily automatable than others. Moreover, in developing KRINO, we also consider the *desirability* of automation and the role of users in relationship to machines as a vital issue. In our view, even if some parts of the sequence of activities involved in argument-checking would be fully automatable, the user should always remain in the lead and the delegation of tasks or subtasks to the machine should never imply loss of control or a shift of responsibility. Nevertheless, the project of (partly) automating the process of argument evaluation has value. Notably, some steps in the normalization of arguments in natural language can be difficult for users with no formal or extensive training in linguistics, pragmatics, or argumentation theory, a task KRINO can assist with. Also, assuming that we are able to build a sufficiently comprehensive and accurate (domain-specific) knowledge base, KRINO can be of great help in assisting users to check the validity of arguments in this respect. We return to our stance about the relations between humans and machines in Section 5.

Here, we further elaborate on the following aspect. We want the user to remain in the lead because we strive to build KRINO as an inspectable, glass-box AI engine that

communicates with the user in natural language. It is not designed as a fully-autonomous engine, but rather as an aid for human agents in the analysis of written text and the disentanglement of critical aspects of the underlying argument structure.

Apart from being designed as an inspectable AI engine, KRINO is also designed to be use-case specific. It may help users carrying out a variety of activities falling under the umbrella of argument-checking, for instance:

1. Checking the logical consistency of technical documents such as software requirement specifications.
2. Assisting the analysis of legal reasoning, for instance by checking consistency between claims and jurisprudence.
3. Assisting doctors to find the correct diagnosis, for instance by checking the consistency between the proposed diagnosis and the available knowledge base, from an argumentative perspective.
4. Identifying fake news and conspiracy theories.
5. Analyzing and assessing arguments put forth in online discussions.

The logic of KRINO is based on the theoretical model combining the linguistic representation framework of Constructive Adpositional Grammars (Gobbo & Benini, 2011) and the argument classification framework of the Periodic Table of Arguments (Wagemans, 2016, 2019, 2020) into an integrated framework for representing linguistic and pragmatic aspects of argumentative discourse (Gobbo et al., 2019). The linguistic part allows parsing a human text into machine structures containing syntactic and semantic information. So the goal of the parsing is not just to recognize particular words and their classes (e.g., nouns, verbs, adjectives, etc.) but also to acquire semantic information carried by the text. Therefore, the parsing algorithm is based on the constructive dictionary which is defined in terms of Constructive Adpositional Grammars. In contrast to traditional dictionaries, the constructive dictionary contains the list of lexemes (i.e., morphemes referring to the real world) and the list of construction rules (including morphology, syntax, and phraseology). Therefore, the parsing is capable of recognizing grammatical aspects (e.g., suffixes, etc.) and through construction rules understand their semantic meaning (e.g. past tense, plural number, etc.). The result of such parsing is then a decomposition of the text into a tree structure containing morphemes and their structural and semantic attributes. The tree structure is then suitable for further processing by high-level algorithms. For instance, if the tree represents an argument, it can be transformed (without changing the meaning) into a normalized form tree intended for the argument evaluation. The constructive dictionary as well as the tree structure are language independent (i.e., the smallest lexical item for KRINO is a morpheme and not a word). Therefore, by providing lexemes and construction rules KRINO can be used with various languages. This, we think, is an important asset of KRINO from the perspective of linguistic justice (Van Parijs, 2011) and epistemic diversity (Gobbo & Russo, 2020).

The approach to argument-checking presented in Section 3 enables KRINO to recognize arguments which can be then evaluated and checked if they are consistent with the knowledge base. Then, using the linguistic part, KRINO can formulate the result (answer) in human language and is also able to provide the explanation of why it came to that conclusion (i.e., why the argument is acceptable or not). The argument evaluation is based on the theory of the Periodic Table of Arguments and the knowledge base (i.e., factual knowledge provided to KRINO). First, by using the knowledge base, it evaluates if the premise and conclusion clauses are true and find the chain of statements proving why they are true. Then it identifies the argument form and uses proving chains of premise and conclusion to find the relation (lever) between the conclusion and the premise. If the relation exists, the argument is evaluated as acceptable. We provide here a very simple example of how KRINO is set to analyze an argument, which will hopefully be useful to readers with relevant background in computer science and closely related fields. The example is taken from https://periodic-table-of-arguments.org/periodic-table-of-arguments/beta-quadrant/argument-from-analogy/, which contains information about its source, an explanation of how to reformulate the natural argument into its canonized form, and an analysis of its basic characteristics. Here, we focus on the way in which KRINO evaluates the argument.

Argument to evaluate:

> Cycling on the grass is prohibited because walking on the grass is prohibited.

Knowledge-base:

```
Walking on the grass damages the grass.
Cycling on the grass damages the grass.
If an activity damages the grass then it is prohibited.
```

KRINO steps:

1. Argument form: a is X because b is X. a (Cycling on the grass) X (is prohibited) because b (walking on the grass) X (is prohibited).
2. Argument lever: Relationship of analogy between a (Cycling on the grass) and b (walking on the grass).
3. Premise clause "`walking on the grass is prohibited`" is true (according to the domain-specific items in the knowledge-base).
4. The lever, i.e. the relationship of analogy between a (Cycling on the grass) and b (walking on the grass) is sound (both damage the grass and that is relevant for being forbidden according to the domain-specific items in the knowledge-base).
5. Conclusion clause "`Cycling on the grass is prohibited`" is true (it is based on a true premise and a sound lever).

The step-wise procedure of this example illustrates how KRINO is able to find links between parts of the argument, if and when there is relevant and appropriate information in the knowledge base. The automatization of argument-checking is, by design, always dependent on some background knowledge that is constructed or validated by the user. This means that KRINO will be able to automatically extract new information from the analyzed text and propose to add items to the knowledge base but these self-learning algorithms remain fully inspectable by the user.

At the time of writing, KRINO is able to analyze simple arguments expressed in natural language, and we expect KRINO to be able to handle more complex arguments, and within a variety of specific contexts, in the near future.

With the aid of KRINO, we aim to make the verification of information, in terms of correctness and completeness of an argument, affordable and accessible to every competent user, resulting in a corroborated belief about analyzed arguments and decision-making. We also provide a tool which gives the user the possibility of enhancing, upgrading, or improving their cognitive environment (and the information they analyze) by making it more transparent, rational, and comprehensible. KRINO complies with standards of transparency because of the principles chosen for developing and designing AI algorithms. KRINO AI algorithms also comply with standards of explainability because they are designed as logic-based. This means that KRINO is designed to be capable of providing users with *reasons* why it came to a certain solution. For instance, combining KRINO with machine learning (neural network) can significantly improve the quality of AI results. KRINO and the results it produces are not an opaque box, and they also crucially depend on the user's choices and domain-specific knowledge at various stages of the process, which includes evaluation and usage of the knowledge base, to be tailored to specific use cases.

To sum up: KRINO designed as an inspectable self-learning AI that will be capable of analyzing arguments in natural language and of forming the knowledge base needed for that purpose. We develop KRINO as an open-source project using the GitHub platform and so all its algorithms are publicly available and inspectable by anyone.

THE SOCIETAL RELEVANCE OF ARGUMENT-CHECKING

In this final section, we explain how the human annotation project of MMGA and the machine annotation project of KRINO can mutually reinforce one another and we discuss what the societal relevance of the combination of the two projects is.

To begin with, the whole project of improving the quality of information via argument-checking is premised on the idea that values such as collegiality or intellectual honesty and humility are the ones we wish to promote (Aberdein & Cohen, 2016; Dalgleish et al., 2017; Kidd, 2016; Tanesini, 2021). With argument-checking, users, agents, and content producers do not act as draconian judges on the mess of online information, but contribute to the quality of information that is shared, distributed, and equally accessible to anyone. With this approach to argument-checking and its automated engine KRINO, we aim to adhere and enhance important ethical considerations. For instance, it is worth distinguishing contexts in which arguments are offered, types of

arguments, and the kind of moral implications that go with them. It matters why and how a given argument is used and it is important not to try to circumvent addressees or hide information.

Also, the project of developing procedures for argument-checking, in both the human and machine annotation variants, aims at improving the digital literacy of users, agents, and content producers, which is an important topic on the digital agenda world-wide. The potential of Human-AI collaboration to combat fake news has already been demonstrated by the project called Demaskuok, which means "debunk" in Lithuanian (see https://www.debunkeu.org/methodology). The AI was developed by the Lithuanian defence ministry in collaboration with Google's Digital News Initiative and Delfi, a media group headquartered in Lithuania's capital, which is able to detect within two minutes of its publication the "patient zeros of fake news" and sends those reports to human specialists for further analysis.

The road ahead of us is steep and we are fully aware of the many challenges faced by both projects. AI, in fact, other than being of potential help in addressing the problem of information quality, may also be a major spreader of fake news (Hao, 2020; Knight, 2021; Lyons, 2020). And so projects like KRINO may be like David in front of Goliath. One challenge of MMGA is that we will never reach enough websites or media platforms or have enough annotators. This is certainly true and this is why, next to projects and initiatives like ours, we also need *systemic* interventions, and these have to be at the level of education, promoted in public spaces and by public institutions. A challenge of KRINO is that it is not intended to be a fully automated AI, and so KRINO users *always* need *some* level of understanding of argumentation theory. KRINO is not a magic bullet to magically turn the internet into a basket of all good pieces of information. It is instead a tool to *help*, where help is needed. These two challenges, together, show the importance of digital literacy, as a necessary component of the education of newer and older generations. But more *literacy* on its own, will not do. Another challenge of KRINO is that, although it is accompanied with a thorough ethics chart, after all it is (and will be) open source. This means that we can't anticipate and prevent all uses of KRINO. What we need is digital literacy *and* cultivation of epistemic and moral virtues in a digital environment. It is high time to reconnect ethics and science and technology in a constructive and productive way. In our view, ethics not a watchdog, or an exercise that happens "after the fact" only (Ratti & Stapleford, 2021; Russo, 2018). We strive to build an ethics stance into our practices, from the set up of training on argument checking to the design of KRINO, specifying, at each and every stage of both these processes, which values guide our practices.

Despite all these challenges, we think MMGA and KRINO projects are worth pursuing. The digital revolution has already happened. It is high time also to move beyond utopian or dystopian attitudes toward technologies (Russo, 2018). What we need more than anything else are projects that believe in the potential of technologies, and that pursue their design and implementations for the common good.

Finally, by combining KRINO and MMGA, we aim to promote a specific normative point of view about the relation between humans and machines, whereby machines

remain at the service of us humans in general, and specifically in this project to improve the quality of online information. We do not buy into the hype of full automation. The question is not posed at the technical but at the normative level. We believe in the value-based *interaction* between humans and machines, and it is in this sense that machines need us more than we do (Russo, 2022). In the footsteps of pioneer of cybernetics Norbert Wiener (1950), we think of technology in general, and AI specifically, as an applied morality, over and above the continuous development and improvement of technical capacities.

REFERENCES

Aberdein, A., & Cohen, D. H. (2016). Introduction: Virtues and arguments. *Topoi*, *35*(2), 339–343. 10.1007/s11245-016-9366-3

Altay, S., Berriche, M., & Acerbi, A. (2021). Misinformation on Misinformation: Conceptual and Methodological Challenges. 10.31234/osf.io/edqc8

Barrera Rodriguez, O. D., Guriev, S. M., Henry, E., & Zhuravskaya, E. (2017). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics*, *182*, 104123. 10.2139/ssrn.3004631

Beaudoin, C. E. (2008). Explaining the relationship between internet use and interpersonal trust: Taking into account motivation and information overload. *Journal of Computer-Mediated Communication*, *13*(3), 550–568. 10.1111/j.1083-6101.2008.00410.x

Borg, S. (2019). We are edging to a world where reality is a matter of personal opinion. *2019*. https://timesofmalta.com/articles/view/we-are-edging-to-a-world-where-reality-is-a-matter-of-personal-opinion.725056

Brave, R. (n.d.). *Post-truth Conference Malta 2019—Talk on Media, Journalism & Fake News*. https://open.spotify.com/episode/3WzhTSRe1TSxnZQKz6e7iN

Brave, R. (2019). Introducing "public annotations" in journalism. *Medium.Com*. Introducing "public annotations" in journalism

Brave, R. (2021). Public rebuttal, reflection and responsibility or an inconvenient answer to fake news. In Grech, A. (Ed.), *Media, technology and education in a post-truth society* (pp. 145–154). Emerald Publishing Limited. 10.1108/978-1-80043-906-120211011

Brave, R., Russo, F., & Wagemans, J. H. M. (2022). Argument-checking: A critical pedagogy approach to digital literacy. *AIUCD 2022 - Digital Cultures. Intersections: Philosophy, Arts, Media. Proceedings of the 11th National Conference, Lecce*, *2022*, 245–248.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*(4), 1073–1095. 10.1007/s11109-019-09533-0

Dalgleish, A., Girard, P., & Davies, M. (2017). Critical thinking, bias and feminist philosophy: Building a better framework through collaboration. *Informal Logic*, *37*(4), 351–369. 10.22329/il.v37i4.4794

De Kerckhove, D. (1998). *Connected intelligence: The arrival of the web society*. London: Kogan Page.

Drutman, L. (2020, June 3). Fact-Checking misinformation can work. But it might not be enough. *FiveThirtyEight*. https://fivethirtyeight.com/features/why-twitters-fact-check-of-trump-might-not-be-enough-to-combat-misinformation/

Ellwart, T., Happ, C., Gurtner, A., & Rack, O. (2015). Managing information overload in virtual teams: Effects of a structured online team adaptation on cognition and performance. *European Journal of Work and Organizational Psychology*, *24*(5), 812–826. 10.1080/1359432X.2014.1000873

Fabry, M. (2017). Here's How the First Fact-Checkers Were Able to Do Their Jobs Before the Internet. Time Magazine. Retrieved July 29, 2022 from https://time.com/4858683/fact-checking-history/

Floridi, L. (2014). *The 4th revolution: How the infosphere is reshaping human reality* (First edition). Oxford University Press.

Floridi, L. (Ed.). (2015). *The onlife manifesto: Being human in a hyperconnected era* (1st ed. 2015). Springer International Publishing: Imprint: Springer. 10.1007/978-3-319-04093-6

Freire, P., Ramos, M. B., & Macedo, D. P. (2014). *Pedagogy of the oppressed: 30th anniversary edition*. United Kingdom: Bloomsbury Publishing.

Gobbo, F., & Benini, M. (2011). *Constructive adpositional grammars: Foundations of constructive linguistics*. Cambridge Scholars Publishing.

Gobbo, F., Benini, M., & Wagemans, J. H. M. (2019). Annotation with adpositional argumentation. *Intelligenza Artificiale*, *13*(2), 155–172. 10.3233/IA-190028

Gobbo, F., & Russo, F. (2020). Epistemic diversity and the question of lingua franca in science and philosophy. *Foundations of Science*, *25*(1), 185–207. 10.1007/s10699-019-09631-6

Graves, L., & Amazeen, M. A. (2019). Fact-Checking as idea and practice in journalism. In L. Graves & M. A. Amazeen (Eds.), *Oxford research encyclopedia of communication*. Oxford University Press. 10.1093/acrefore/9780190228613.013.808

Hao, K. (2020, August 14). A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it. *MIT Technology Review*. https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/

Hinton, M., & Wagemans, J. H. M. (2022). Evaluating reasoning in natural arguments: A procedural approach. *Argumentation*, *36*(1), 61–84. https://doi.org/10.1007/s10503-021-09555-1

Islam, N., Shaikh, A., Qaiser, A., Asiri, Y., Almakdi, S., Sulaiman, A., Moazzam, V., & Babar, S. A. (2021). Ternion: An autonomous model for fake news detection. *Applied Sciences*, *11*(19), 9292. 10.3390/app11199292

Jones, Q., Ravid, G., & Rafaeli, S. (2004). Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information Systems Research*, *15*(2), 194–210. 10.1287/isre.1040.0023

Kidd, I. J. (2016). Intellectual humility, confidence, and argumentation. *Topoi*, *35*(2), 395–402. 10.1007/s11245-015-9324-5

Knight, J., Dooly, M., & Barberà, E. (2020). Getting smart: Towards critical digital literacy pedagogies. *Social Semiotics*, 1–24. 10.1080/10350330.2020.1836815

Knight, W. (2021, May 24). AI Can write disinformation now—and dupe human readers. *Wired*. https://www.wired.com/story/ai-write-disinformation-dupe-human-readers/#:~:text=Over%20six%20months%2C%20a%20group,on%20particular%20points%20of%20disinformation

Li, C.-Y. (2017). Why do online consumers experience information overload? An extension of communication theory. *Journal of Information Science*, *43*(6), 835–851. 10.1177/0165551516670096

Lévy, P. (1997). *Collective intelligence: Mankind's emerging world in cyberspace*. Plenum Trade: New York.

Lyons, K. (2020, August 16). A college student used GPT-3 to write fake blog posts and ended up at the top of Hacker News. *The Verge*. https://www.theverge.com/2020/8/16/21371049/gpt3-hacker-news-ai-blog

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. Data & Society. https://datasociety.net/wp-content/uploads/2017/05/DataAndSociety_MediaManipulationAndDisinformationOnline-1.pdf

MMGA. (2019, March 13). Introducing "public annotations" in journalism. *Medium.Com*. https://medium.com/@MakeMediaGreatAgain/introducing-public-annotations-in-journalism-e688b04be903

Nieman, C. (2020, November 27). 'Whoever does not study rhetoric will be a victim of it'. From fact-checking to argument-checking as Award Nominated researchers of University of Amsterdam join MMGA with human-AI framework. *Internet Society Netherlands*. https://isoc.nl/nieuws/whoever-does-not-study-rhetoric-will-be-a-victim-of-it/

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, *118*(15), e1912440117. 10.1073/pnas.1912440117

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. 10.1007/s11109-010-9112-2

Nyhan, B., & Reifler, J. (2012). *Misinformation and Fact-checking: Research Findings from Social Science* (New America). New America Foundation. https://cpb-us-e1.wpmucdn.com/sites.dartmouth.edu/dist/5/2293/files/2021/03/Misinformation_and_Fact-checking.pdf

Palfrey, J., & Gasser, U. (2016). *Born digital: How children grow up in a digital age*. Basic Books. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=4785961

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. 10.1037/xge0000465

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), e2104235118. 10.1073/pnas.2104235118

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. 10.1016/j.tics.2021.02.007

Plug, H. J., & Wagemans, J. H. M. (2020). From fact-checking to rhetoric-checking: Extending methods for evaluating populist discourse. In van der Geest, I., Jansen, H., & van Klink, B. (Eds.), *Vox Populi* (pp. 236–252). Edward Elgar Publishing. 10.4337/9781789901412.00023

Ratti, E., & Stapleford, T. A. (Eds.). (2021). *Science, technology, and virtues: Contemporary perspectives* (1st ed.). Oxford University Press. 10.1093/oso/9780190081713.001.0001

Roetzel, P. G. (2019). Information overload in the information age: A review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research*, *12*(2), 479–522. 10.1007/s40685-018-0069-z

Russo, F. (2018). Digital technologies, ethical questions, and the need of an informational framework. *Philosophy & Technology*, *31*(4), 655–667. 10.1007/s13347-018-0326-2

Russo, F. (2022). *Techno-scientific practices. An informational approach*. Rowman & Littlefield International.

Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*(4), 713–724. 10.1086/426605

Swar, B., Hameed, T., & Reychav, I. (2017). Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. *Computers in Human Behavior*, *70*, 416–425. 10.1016/j.chb.2016.12.068

Tanesini, A. (2021). Virtues and vices in public and political debates. In Hannon, M. & de Ridder, J. (Eds.), *The Routledge handbook of political epistemology*. Routledge/Taylor & Francis Group.

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*(3), 460–480. 10.1080/10584609.2015.1102187

Uscinski, J. E., & Butler, R. W. (2013). The epistemology of fact checking. *Critical Review*, *25*(2), 162–180. 10.1080/08913811.2013.843872

Van Parijs, P. (2011). *Linguistic justice for Europe and for the world*. Oxford University Press. 10.1093/acprof:osobl/9780199208876.001.0001

Wagemans, J. H. M. (2016). Constructing a periodic table of arguments. *SSRN Electronic Journal*. 10.2139/ssrn.2769833

Wagemans, J. H. M. (2019). Four basic argument forms. *Research in Language*, *17*(1), 57–69. 10.2478/rela-2019-0005

Wagemans, J. H. M. (2020). Why missing premises can be missed: Evaluating arguments by determining their lever. In Cook, J. (Ed.), *Proceedings of OSSA 12: Evidence, persuasion & diversity*. OSSA Conference Archive. https://scholar.uwindsor.ca/ossaarchive/OSSA12/Saturday/1

Wagemans, J. H. M. (2021). *Argument type identification procedure (ATIP) – Version 3*. www.periodic-table-of-arguments.org/argument-type-identification-procedure

Wagemans, J. H. M. (Forthcoming). On the hermeneutics of persuasive discourse: How to identify an argument type? *Journal of Pragmatics*.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-Checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. 10.1080/10584609.2019.1668894

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (No. 162317GBR; Éditions Du Conseil de l'Europe). https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

White, M., & Dorman, S. M. (2000). Confronting information overload. *Journal of School Health*, *70*(4), 160. Gale Academic OneFile. https://link.gale.com/apps/doc/A61995006/AONE?u=anon~25ba5948&sid=googleScholar&xid=a2a7c170

Wiener, N. (1950). *The human use of human beings* ((1989)). Free Association Books.

Wintersieck, A. L. (2017). Debating the Truth: The impact of fact-checking during electoral debates. *American Politics Research*, *45*(2), 304–331. 10.1177/1532673X16686555

Zandbergen, A. D. (2011). *New edge: Technology and spirituality in the San Francisco Bay Area* [Universiteit Leiden].