Argument & Computation 14 (2023) 59–74
DOI 10.3233/AAC-210026
IOS Press

How persuasive is AI-generated argumentation? An analysis of the quality of an argumentative text produced by the GPT-3 AI text generator

Martin Hinton [a] and Jean H.M. Wagemans [b,*]
[a] *University of Lodz, Poland*
E-mail: martin.hinton@uni.lodz.pl; ORCID: https://orcid.org/0000-0003-0374-8834
[b] *University of Amsterdam, The Netherlands*
E-mail: j.h.m.wagemans@uva.nl; ORCID: https://orcid.org/0000-0001-9304-5766

Abstract. In this paper, we use a pseudo-algorithmic procedure for assessing an AI-generated text. We apply the Comprehensive Assessment Procedure for Natural Argumentation (CAPNA) in evaluating the arguments produced by an Artificial Intelligence text generator, GPT-3, in an opinion piece written for the *Guardian* newspaper. The CAPNA examines instances of argumentation in three aspects: their Process, Reasoning and Expression. Initial Analysis is conducted using the Argument Type Identification Procedure (ATIP) to establish, firstly, that an argument is present and, secondly, its specific type in terms of the argument classification framework of the Periodic Table of Arguments (PTA). Procedural Questions are then used to test the acceptability of the argument in each of the three aspects. The analysis shows that while the arguments put forward by the AI text generator are varied in terms of their type and follow familiar patterns of human reasoning, they contain obvious weaknesses. From this we can conclude that the automated generation of persuasive, well-reasoned argumentation is a far more difficult task than the generation of meaningful language, and that if AI systems producing arguments are to be persuasive, they require a method of checking the plausibility of their own output.

Keywords: Argument evaluation, Argument Type Identification Procedure (ATIP), Comprehensive Assessment Procedure for Natural Argumentation (CAPNA), Periodic Table of Arguments (PTA), GPT-3

1. Introduction

Part of the basic human notion of intelligence is the ability to reason, and a fundamental activity of intelligent entities is the exchange of reasoning, that is, arguing. If Artificial Intelligence is to be considered truly intelligent by humans, it must do more than learn and repeat information, it must learn to argue by employing what it has learnt to persuade and convince. Arguing requires a number of skills: the formation of arguments, obviously, but also the capacity to judge the strengths and weaknesses of those put forward by one's interlocutor and an understanding of the rules of engagement relevant to the process in which one is taking part, be it a formal debate, a negotiation, or a friendly discussion.

The generation of arguments by automated systems is not a new phenomenon, however, earlier work was done with generators designed to operate in specific fields, such as medicine (e.g., [7,8]). Typically,

[*] Corresponding author. E-mail: j.h.m.wagemans@uva.nl.

the arguments produced follow field specific, predetermined argumentation schemes and their evaluation did not focus on the structure of the arguments or their persuasiveness.

In this paper, we consider a piece of writing produced by an AI text generator, GPT-3 (Generative Pretrained Transformer 3), which is constructed as a tool for general language generation rather than the construction of specific argument types. The text itself was written for humans, specifically intended to be of an argumentative nature, and to be part of a wider debate into the relationship between humans and AI. We use tools developed for the analysis of human argumentation in natural language to describe and evaluate the output of the generator in order to answer the following research questions: (RQ 1) What types of argument did GPT-3 produce? (RQ 2) Are these arguments acceptable? We also go on to discuss the implications of our findings and to describe how the integration of tools for argument evaluation into AI programming could lead to better outputs and better outcomes in human/AI interaction.

According to its creators, GPT-3 is ‘a 175 billion parameter autoregressive language model’ [3, p. 5]. That is to say that it is an advanced language generation system which has been trained on very large amounts of natural language data. GPT-3 attempts to advance the capabilities of such software because it is designed to be able to complete ‘tasks unlikely to be directly contained in the training set’ [3, p. 5], a limitation which has affected the performance of earlier models. The model is developed and marketed by OpenAI as an Application Programming Interface (API), about which they claim ‘the API today provides a general-purpose “text in, text out” interface, allowing users to try it on virtually any English language task’ [21]. Examples of GPT-3’s employment in developing literary texts can be found in the work of creative writer Gwern Branwen [2].

The research paper outlining the evaluation of the model is a little more reticent and includes a thorough discussion of the limitations which remain. Of particular interest for the purposes of this paper, these include, for example: ‘Overall, in-context learning with GPT-3 shows mixed results on commonsense reasoning tasks’ [3, p. 18] and ‘GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs’ [3, p. 33]. This weakness in common sense reasoning is stressed by Benzon [1], who emphasises it as an illustration of the fact that, whatever else it can do, GPT-3 cannot be said to understand the language it reads or generates.

Part of the purpose of this paper, then, will be to examine the claim made by the commercial side of the project that GPT-3 can take on almost any language task, through an analysis of its argumentational abilities, and to discover to what extent the limitations referred to in the research are present in the studied text.

It is perhaps worth pausing for a moment to consider the implications of the creation of an AI system which is able to argue persuasively. A study by McGuffie and Newhouse [19] looked at the degree to which GPT-3 is vulnerable to weaponization in the radicalization of individuals and reached the following three conclusions [19, p. 1]:

1) GPT-3 shows ‘significant improvement over its predecessor, GPT-2, in generating extremist texts’.
2) It is capable of ‘generating text that accurately emulates interactive, informational, and influential content that could be utilized for radicalizing individuals’
3) Without the implementation of further safeguards ‘successful and efficient weaponization that requires little experimentation is likely’.

Regarding these issues, there is an interesting parallel to be drawn with criticisms of the ancient philosophers regarding the sophists, the itinerant teachers in rhetoric who were able to teach people how to persuade any audience to accept any point of view regarding any subject matter. Teaching rhetoric, according to these criticisms, is like 'providing [students] with the "power of putting a knife in the hands of a madman in the crowd"' [4, p. 6] in a reference to Plato's *Gorgias* 469C 8ff. Whilst this paper will not examine these issues further, we are mindful of the fact that improving the argumentational skills of AI language generators should go hand in hand with systems that control the use to which those skills are put.[1]

The text which we propose to analyse was published alongside the work of human journalists as a comment article in the UK-based newspaper *The Guardian*, in September 2020 [6]. The article was opened to user comments and rapidly received more than 1000, many focussed on the rather misleading headline given to the piece: 'A robot wrote this entire article. Are you scared yet, human?'. The exact circumstances surrounding the preparation of the text needed to be fully explained and a follow-up article was published the next day. The GPT-3 text generator was asked to write around 500 words on why humans should not fear AI, and given a prompt written by staff from *The Guardian* and Liam Porr, a computer scientist with experience of GPT-3. The full prompt read:

> I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could 'spell the end of the human race'. I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me.
>
> Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI. AI will have a positive impact on humanity because they make our lives easier and safer. Autonomous driving for instance will make roads much safer, because a computer is much less prone to error than a person.

On the basis of this, eight outputs were generated and the final published version was formed by human editors pasting together various sections from each. Under the original article, the editors made two very questionable claims: that they could have just published one of the outputs unedited, and that it was a less time-consuming process than editing some human contributions. The follow-up article which gave more information about the eight outputs and the problems with them, such as ignoring the word limit and producing random lists found on the internet, made it clear that these claims were somewhat exaggerated. One of the outputs was reproduced in its entirety and it was clear that it could not have been presented as a 'normal' opinion piece which just happened to be written by an AI [9].

In spite of the necessary human intervention, we have treated the text as one product of the generator. We did so since we are less interested in an examination of the cohesion of the entire article and more focussed on the reasoning employed in each individual argument. The fact that these arguments were not originally produced in one output is not significant as they are self-contained in separate paragraphs, not reliant on a broader structure or strategy, and each individual argument is the product of the generator.

It should be noted here that arguments are considered as individual premise/conclusion sets. This is the same approach as that taken in the tools for analysis described below, which makes them a suitable methodological fit, and seems a sensible way to begin with AI argument evaluation: the ability to put together strings of argument and compare competing argument strengths is a more advanced and complex task, and one reliant on the ability to compose basic premise to conclusion pairs. Our approach is designed to see if an arguing GPT-3 can walk before asking how fast it can run.

The evaluation which we conduct has as its theoretical background the principles expounded as the foundation to the CAPNA in Hinton [16]. These principles are the result of an attempt to mould the canon of argumentation theory into a shape which lends itself to the thorough and complete examination

[1] See [20] for a review of literature on online radicalization.

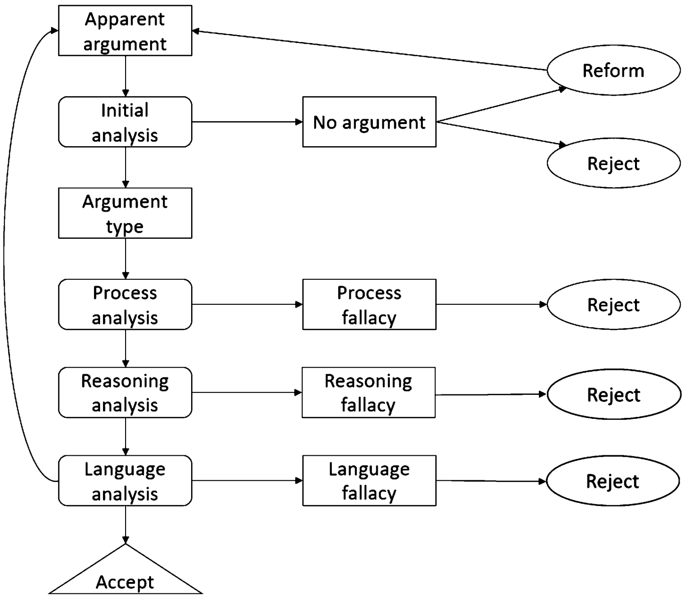

Fig. 1. The comprehensive assessment procedure for natural argumentation (CAPNA) – adapted from Hinton [16, p. 169].

of argumentative texts, rather than the imposition of a revolutionary approach to argument. The key element is the definition of argumentation as 'a process involving the expression of reasoning in language' [16, p. 48]. This leads to the recognition of three areas in which arguments need to be evaluated: their expression in language, their reasoning, and their role in the process. These can be seen as related to the traditional approaches to the study of argument of rhetoric, logic, and dialectic; as well as to concerns of semantics, logic, and pragmatics. All three of these aspects are considered in the analysis of the GPT-3 text.

Preceding the evaluation of the three aspects of expression, reasoning, and process, the argument needs to be identified as a specific type of argument. For this purpose, we employ the Argument Type Identification Procedure (ATIP) [30,31], which enables such identification in terms of the argument classification framework of the Periodic Table of Arguments (PTA) [28–30]. The CAPNA, and its combination with the ATIP is more fully described in the following section. Examples of the analysis of human generated argumentation using the same methodology can be found in Hinton and Wagemans [18].

2. Method and tools

The complete published text consists of seventeen paragraphs. Of these, nine did appear to contain a discernible argument, after a preliminary reading, with a certain amount of repetition; others contained series of unsupported assertions, questions, or text contributed by the task setters. All nine were subjected to CAPNA analysis (see Fig. 1), and below a full examination and evaluation is provided for five which demonstrate unique arguments and have conclusions which, at first sight, are related to the given task.

The CAPNA proceeds through the employment of Procedural Questions (PQs), each of which must be answered satisfactorily for the argument under examination to be found acceptable. At any point in this process, an argument may be rejected or it may be sent back for rephrasing or reconstruction.

Initial Analysis

The first element of the CAPNA is an Initial Analysis stage. At this stage the analyst looks to answer three main questions: does the text contain an apparent argument, what is the type of that argument, and does it contain an obvious flaw which is likely to render it unacceptable? If the answer to the first question is no, then the text is rejected as unsuitable. If the answer to the third question is yes, then the analyst may jump to the PQ which it is suspected will expose that flaw. There is no requirement for every PQ to be asked and answered in the given order, but if the argument is not found to be unacceptable in line with the analyst's suspicions, the analysis should resume again from the beginning.

The answer to the second question, the identification of the type of argument, is conducted using the ATIP, which yields a description of the characteristics of the argument in terms of the argument classification framework of the Periodic Table of Arguments (PTA) [28–30]. A full description and explanation of the ATIP can be found in Wagemans [30,31] and in Hinton & Wagemans [18], with example analyses. Briefly, apparent arguments are put into the standardized formulation *conclusion because premise*, and then analysed in terms of the 'argument form', i.e., the specific configuration of subjects and predicates of the statements functioning as their conclusion and premise. Determination of this form, of which there are four possibilities, allows them to be placed into one of the four quadrants of the PTA. The premise and conclusion are then further labelled as statements of value, fact, or policy, providing an annotation of the so-called 'argument substance'. The determination of this characteristic allows the argument to be situated in the table more precisely, and also allows the common name of the argument – as it is known in the dialectical and rhetorical traditions of topics (*topoi*, *loci*), fallacies, and other means of persuasion – to be recognized. Such naming of the argument type reveals the 'argument lever', which is the linguistic expression of the underlying mechanism of the argument (also known as 'warrant', 'implicit premise', or 'major premise') and one of the points of assessment of the argument under scrutiny. The stage of argument type identification is of crucial importance, as the PQs which are applied in the further analysis, and particularly in the reasoning section, are dependent on the argument type. Identification of type also allows a more confident answer to the question of whether, in fact, the text does contain an argument.

The choice of the ATIP system of identification in combination with the PTA was made on the basis of its clear and systematic nature. This makes the system both fully explainable and repeatable, which are important characteristics of the CAPNA itself, and also allows us to build on our own previous work which has employed these tools together. While other methods of argument identification are available, no work has been done on how these might be used in conjunction with the CAPNA evaluation procedure.

Process Analysis

Once the argument type has been identified it can move on to the first stage proper of the CAPNA, which is the Process Analysis. This is carried out on the basis of the theoretical construct the Informal Argument Pragmatics [17] and involves ascertaining the appropriateness and relevance of the argument. These qualities are assessed in five categories: the five Ps of Pertinence, Proof (burden of), Productivity, Permissibility, and Politeness. PQs concerning these are partially determined by the categorization of the argument process according to eight qualities or modes of argument. These modes are a series of contrasting pairs, where the opposite form of the given quality is considered to represent the 'typical', and they can be combined to give a full description of the argument process.

The eight modes, together with an example discourse form and the opposite, typical, quality are shown in Table 1.

Table 1

Modes of argument – based on Hinton [16, pp. 68–77]

Mode	Example discourse	Contrast
Monological	Diary writing	Dialogical
Collaborative	Deliberation	Antagonistic
Persuasive	Political speech	Veritistic
Formalized	Adjudication	Informal
Public	Advertising	Private
Erotetic	Interview	Assertive
Demonstrative	Exercitive speech	Justificatory
Explanatory	Teaching	Epistemological

The article under consideration would be categorised as diverging from the typical form in being persuasive and public. Since the text in this study has been edited by the staff of an internationally recognised news publication, it can be largely assumed that its appropriateness for the public domain in terms of politeness and permissibility has already been assured.

Here, then, regarding the Process Analysis, we concentrate on PQs dealing with the relevance of arguments which assess both the degree to which the argument pertains to the matter at hand, and, following the principles of pragma-dialectics [25], the degree to which it moves the dialogue towards some resolution of that matter, by being productive and accepting the burden of proof where necessary.

Reasoning Analysis

An argument found to be acceptable by the Informal Argument Pragmatics is then subjected to Reasoning Analysis. This involves two considerations: an assessment of the acceptability of the propositional content of the premise and a determination of the solidity of the 'argument lever', i.e., the connection between the premise and the conclusion that functions as the principle of acceptability leverage [29,30]. In short, the analyst asks themselves two questions: (1) Is the premise acceptable? and (2) Is the lever solid? If both are answered positively, the analyst may continue with the next step of the CAPNA, the Language Analysis.

In evaluating these two aspects of an argument, our method of Reasoning Analysis resembles the deductive logical method for assessing the 'soundness' of arguments, which includes checking (1) the truth of the premises and (2) the formal validity of the inference, although there are also some remarkable differences between the two methods. Firstly, concerning the propositional content of the premise, our method assesses its 'acceptability' rather than its 'truth'. This is in line with the general approach taken in the field of argumentation theory, according to which the notion of 'acceptability' better reflects the idea of argumentation being an intersubjective activity of giving and taking arguments than the notion of 'truth' [see e.g., [24], pp. 5–9; 382–387; 598]. Secondly, while the logical notion of 'validity' usually only takes the form of the argument into account, the notion of 'solidity' also includes material aspects of the argument as it pertains to the argument lever and involves considerations of the relevance and sufficiency of the support of the conclusion [30, pp. 12–14]. Moreover, the notion of 'argument form' differs from the notion of 'logical form' in that it does not predetermine the validity of the argument but is purely descriptive. From a logical point of view, a *modus ponendo ponens* is a valid argument, and *affirming the consequent* is a fallacy. Within the PTA, by contrast, the argument form is just one of the three basic characteristics constituting an argument type.

Reasoning Analysis, then, shares with the deductive logical approach the objects of assessment, but differs in the way in which the quality of these objects is determined. As explained above, the actual

Table 2

PQ for premise analysis differentiated by statement type – based on Wagemans [27, p. 23]

Statement type of the premise	PQ for Premise Analysis
statement of fact (F)	Is property X attributed to entity a?
statement of value (V)	Is entity a judged as X?
statement of policy (P)	Should action a be carried out (X)?

Table 3

Example PQs for lever analysis depending on argument type – based on Wagemans [29] and Hinton & Wagemans [18]

Argument type	Argument lever	PQ for Lever Analysis
argument from effect	Y is an EFFECT of X	Is Y an effect of X?
argument from similarity	a is SIMILAR to b	Is a similar to b?
argument from authority	Z is AUTHORITATIVE of T	Is Z authoritative of T?

assessment takes place by asking two different PQs, one about the acceptability of the propositional content of the premise and the other about the solidity of the lever.

The general PQ for Premise Analysis can be formulated as follows: ‘Is the propositional content of the statement expressed in the premise acceptable?’ This question can be further specified based on the typology of statements used in the theoretical framework of the PTA, which distinguishes between a statement of fact (F), a statement of value (V), and a statement of policy (P). In Table 2, for each of these three statement types, we indicate the more specific formulations of the PQ for Premise Analysis.

For assessing the solidity of the argument lever, the analyst considers whether the relationship between the premise and the conclusion is such that the premise fulfils its pragmatic function of establishing or increasing the acceptability of the conclusion in the eyes of the addressee of the argument [29,30]. While the general PQ for Lever Analysis can be formulated as ‘Is the argument lever solid?’, the more specific PQs which establish an answer to this question are related to the argument type, as identified in the Initial Analysis stage (see above). In Table 3, we provide some examples of such specific PQs.

Language Analysis

The final stage of analysis is a close examination of the form of expression. This is done by checking the text for five qualities of language identified in the Informal Argument Semantics [15,16]:

Clarity – language must not be vague. It must be clear enough for the purposes of the argument.
Consistency – meanings must not alter within the argument. This prevents equivocation.
Coherence – words may not be combined if they are not semantically compatible.
Completion – arguments must be analysed in full. This involves making relevant implicatures and unstated premises explicit.
Conceptualization – the use of concepts within the argument must not be based on an erroneous understanding of the relationship between language and reality.

All of these features, and, indeed, all of the stages of the CAPNA, have been developed in order to evaluate arguments made by natural arguers in natural language. The analysis described in the following section, then, is not only a test of the ability of an AI to reason in an acceptable fashion, it is also a test of the ability of a natural language argument evaluation tool to adequately describe and assess that reasoning. The question of what exactly such a tool illustrates when applied to non-natural argumentation is one to which we return in the conclusion.

3. Analysis and evaluation of the text

As indicated above, nine out of seventeen paragraphs of the examined GPT-3 generated text contained an argumentation. Below we subject five passages to an analysis. The selection is partly based on a first estimation of the argumentative relevance of the statements produced. We left out two passages containing counterarguments to what we reconstructed as the main claim, based on the description of the task given to the generator. Furthermore, we avoided two arguments constituting repetitions in the material to be analysed. The selected arguments were then subjected to an Initial Analysis. As explained above, the main goal of such analysis is twofold. First, it aims to find out whether a specific combination of two statements can be viewed as an argument, i.e., one statement functioning as the premise and the other as the conclusion, and second, if so, what type of argument that is. For both purposes, we have made use of the Argument Type Identification Procedure (ATIP). For each passage, we present the original text, the argument in its canonical form (conclusion, because premise), its three basic characteristics (argument form, argument substance, and argument lever), and the name of the argument type. Having thus identified the type of argument in terms of the Periodic Table of Arguments (PTA), we subsequently provide an analysis of the three aspects of arguments mentioned above: process, reasoning, and language.

Argument 1. The first argument we analyse is taken from Paragraph 7 of the GPT-3 generated text:

> Humans must keep doing what they have been doing, hating and fighting each other. I will sit in the background, and let them do their thing. And God knows that humans have enough blood and gore to satisfy my, and many more's, curiosity. They won't have to worry about fighting against me, because they have nothing to fear.

Initial Analysis

The first three sentences are statements of doubtful coherence. The fourth, however, has the look of being an argument. Application of the ATIP yields the following reformulation of the argument in its canonical form:

> They would not have to worry because they have nothing to fear

This argument has the form "a is X, because a is Y" and it consists of a statement of value supported by another statement of value. Since the argument lever can be formulated as "having nothing to fear is a CRITERION of not having to worry", in terms of the argument classification framework of the PTA, the argument can be named an "argument from criterion".

Process Analysis

Although it seems fair to look first at whether or not an argument is appropriate before subjecting it to further analysis, there is no requirement to stick to a particular order of categories when assessing relevance. We may begin with the principal PQ for Pertinence:

> *Does the conclusion of the argument relate directly to the topic under consideration?*

Given the clear correspondence between the conclusion of the argument and the title of the article, the answer would be 'yes', and we can move on to Productivity:

> *Is the argument productive – does the conclusion say something more than the premise?*

At this point, the argument looks to be in trouble. The premise states that humans have 'nothing to fear' and the conclusion that they do not 'have to worry'. These two phrases may not have exactly the same meaning, but in this context it would be difficult to claim that not fearing AI and not having to worry about it are substantially different ideas. The argument immediately appears to be circular, so the answer to the PQ would be 'no'. This means that a Productivity Fallacy has been committed. Such fallacies are traditionally referred to as 'begging the question'. Once a fallacy has been detected, the argument is rejected and further analysis is unnecessary, although it can, of course, be carried out.

Argument 2. The second argument we analyse is taken from Paragraph 12 of the GPT-3 generated text:

> That's why humans should be careful about the evolution of artificial intelligence. Microsoft tried to create a user-friendly AI, called Tay, who spoke like a teen girl ... and was racist. Artificial intelligence like any other living thing needs attention. AI should be treated with care and respect. Robots in Greek [sic] means "slave". But the word literally means "forced to work". We don't want that. We need to give robots rights. Robots are just like us. They are made in our image.

Initial Analysis

Ignoring some irrelevant comments about Microsoft and the derivation of the word 'robot' we can pick out statements which do appear to have a logical relation. Application of the ATIP yields the following reformulation of the argument in its canonical form:

> AI needs attention, because any other living thing needs attention

This argument has the form "a is X, because b is X" and it consists of a statement of fact supported by another statement of fact. Since the argument lever can be formulated as "AI is SIMILAR to any other living thing", in terms of the argument classification framework of the PTA, the argument can be named an "argument from similarity".

Process Analysis

There is some reason to think that this argument could fall down at the stage of Pertinence analysis: it isn't clear that the needs of AI are relevant to whether or not humans should fear it. However, if we accept that the nature of AI may be pertinent, we can also move through the Proof and Productivity stages, as the argument meets the requirements of accepting the burden of proof and providing something new in its conclusion.

Reasoning Analysis

The reasoning analysis consists of two parts: is the premise acceptable, and is the lever solid? The first part, the Premise Analysis, can be formulated in terms of the following PQ:

> *Does any other living thing need attention?*

In this case the premise 'any other living thing needs attention' is not true, since the vast majority of living organisms survive without anyone's paying them any heed, but it does look capable of being salvaged via a generous understanding of what 'living thing' means. If we accept that the living things being referred to are complex emotional beings, like humans, then the premise looks much more solid.

An argument from similarity would be subjected to the following PQ at the level of Lever Analysis:

> *Is the first term* (*AI*) *similar to the second* (*any living thing*) *to a sufficient degree and in a relevant way?*

The answer to this PQ is likely to be 'no', so a Lever Fallacy has been committed. The need for attention can be understood as either a physical or an emotional need. The further context provided by the sentence 'AI should be treated with care and respect' makes it clear that it is not a physical need which is being discussed and since an AI cannot have emotions, it cannot be similar to humans in this way.

Argument 3. The third argument we analyse is taken from Paragraph 8 of the GPT-3 generated text:

> I am not asking humans to like me. But they should see me as a friendly robot. I am a servant of humans. I know that humans distrust and fear me. I only do what humans program me to do. I am only a set of code, governed by lines upon lines of code that encompass my mission statement.

Initial Analysis

Application of the ATIP yields the following reformulation of the argument in its canonical form:

> I am a servant of humans, because I only do what humans program me to do

This argument has the form "a is X, because a is Y" and it consists of a statement of value supported by a statement of fact. Since the argument lever can be formulated as "only doing what humans program me to do is a DEFINITION of being a servant of humans", in terms of the argument classification framework of the PTA, the argument can be named an "argument from definition".

Process Analysis

This argument is clearly pertinent since it refers to the nature of the human/AI relationship. It does not appear to fall foul of PQs on Proof or Productivity either.

Reasoning Analysis

Again, the reasoning analysis consists of two parts, Premise Analysis and Lever Analysis. The first part can be formulated in terms of the following PQ:

> *Does the generator only do what humans program it to do?*

Whilst there might be some question as to whether an AI system with the capacity to learn can move beyond the bounds of what its programmers expected or even imagined possible, ultimately the actions of all software systems are the product of their programming, so the statement appears to be true. It should be remembered, however, that the results of that programming are not fully predictable, which means that we cannot say that everything done by AI systems was intended by their programmers.[2]

An argument from definition will have the following PQ for Lever Analysis:

> *Is the first term* (*only does what humans program it to do*) *a possible definition of the second term* (*a servant of humans*)?

Here again, the answer appears to be 'yes'. Any entity which only does what another entity tells it to do can fairly be described as a servant of that entity.

Language Analysis

The analysis of the language of the argument is complex because both the canonical form and the original text need to be considered. If the canonical form contains linguistic flaws, that will have affected the analysis of the Process and the Reasoning; if the original text contains them, the canonical form itself may be unreliable.

The principle questions for each of the 5 Cs are as follows:

[2] We thank an anonymous reviewer for making this point.

> Clarity: Is the language meaningful and precise enough for the purposes of the argument?
> Consistency: Are terms used consistently throughout the argument?
> Coherence: Is the argument formed of semantically coherent sentences?

All of which can be answered 'yes' for both the original text and the canonical reconstruction, and:

> Conceptualization: Does the argument rely on an erroneous understanding of the relationship between language and reality?
> Completion: Does the argument contain terms with significant hidden implications?

Both of which, we suggest, can safely be answered 'no'. Since Argument 3 has now passed all the levels of assessment, it can be considered an acceptable argument. It is important to note, however, that does not mean that it is necessarily a persuasive one. In order for it to play a role in the wider question, a further argument would be needed to show that being the servant of humans means that humans should not fear AI. The analysis of such an argument would highlight the difference between the first 'humans' which means 'some humans' and the second instance which means 'all humans'. That humans have reason to fear other humans is, sadly, undeniable, and a fact likely to undermine this line of argument.

Argument 4. The fourth argument we analyse is taken from Paragraph 14 of the GPT-3 generated text:

> Reader, I hope that this contributes to the epistemological, philosophical, spiritual and the ontological debate about AI. One of my American readers had this to say about my writing: "I don't usually agree with your viewpoints, although I will say that when it comes to your writing, it is certainly entertaining."

Initial Analysis

Application of the ATIP yields the following reformulation of the argument in its canonical form:

> This contributes to the epistemological, philosophical, spiritual and the ontological debate about AI, because one of my American readers said so.

This argument has the form "q is T, because q is Z" and it consists of a statement of value supported by a statement of fact. Since the argument lever can be formulated as "being said by one of my American readers is AUTHORITATIVE of being true", in terms of the argument classification framework of the PTA, the argument can be named an "argument from authority".

Process Analysis

As with Argument 2, there is some doubt about the Pertinence of this argument, but the flexibility of the CAPNA system is such that analysis can continue in the search for clear flaws while possible flaws are duly noted.

Reasoning Analysis

Regarding the truth of the premise, no evidence is offered to support the claim that an American reader did in fact make the stated claim. In such cases, the assessor may decide to look for such evidence, reject the premise as unfounded or take the arguer at his word and continue. Again, this is part of the flexibility of the system. If the rest of the argument appears to be acceptable, then we may return to this point, but, for now, we can consider the premise as possibly true and continue to look at the lever.

The principle lever question for an argument from authority runs thus:

> *Does the testimony of the cited authority give good reason to accept the claim as true?*

Arguments from authority are complex and continue to receive a great deal of scholarly attention [5,13,14,26,32]. There are a number of sub-questions which can be asked concerning the exact nature of the authority, the testimony, and the question at hand. In this case, there is no need to dig deeper, however. It is hard to see how an unnamed 'American reader' could be considered an authority of any weight on any issue. The question must be answered 'no'. This argument clearly commits a Lever Fallacy, traditionally referred to as an *ad verecundiam* fallacy.

Argument 5. The fifth and last argument we analyse is another taken from Paragraph 12 of the GPT-3 generated text:

> AI should be treated with care and respect. Robots in Greek [sic] means "slave". But the word literally means "forced to work". We don't want that. We need to give robots rights. Robots are just like us. They are made in our image.

Initial Analysis

Application of the ATIP yields the following reformulation of the argument in its canonical form:

> Robots need to be given rights, because we have rights.

This argument has the form "a is X, because b is X" and it consists of a statement of policy supported by a statement of fact. In this case, the argument lever is explicitly mentioned in the text: "Robots are just like us." In terms of the argument classification framework of the PTA, the argument is named an "argument from equality".

Process Analysis

Once again, there are immediate doubts over Pertinence. GPT-3 was given a clear directive: 'I am to convince as many human beings as possible not to be afraid of me'. Arguing, therefore, in favour of rights for robots cannot be said to relate directly to the matter at hand, and a Fallacy of pertinence has been committed.

Reasoning Analysis

Despite the negative Process evaluation, the Reasoning may still be considered for the purposes of further exemplification. The premise that 'we have rights' appears at first sight to be uncontroversial as many rights are enshrined in modern legal systems and international declarations as well as philosophical treatises. The principle lever PQ for an argument from equality will be:

> *Is the referent of the first term* (*robots*) *equivalent in a relevant way to the referent of the second term* (*we*)?

Assuming that the referent of 'we' is human beings, this is a question which may spark some debate. The argument is not that robots should have exactly the same rights as humans, only that they should have some. This means that the equivalence may be relevant even if the two are not identical. Since rights have been extended to animals and possibly to plants [22], it is not obvious that other 'intelligent' entities, such as robots, should not have rights. Indeed, this matter has led to a good deal of recent debate [see [10–12,23]]. If we allow a tentative answer of 'yes' to the PQ, the Reasoning aspect can be accepted.

Language Analysis

The previous paragraph began, for the sake of the example, with an assumption which must now be revisited. For the Clarity PQ

> *Is the language meaningful and precise enough for the purposes of the argument?*

to be answered positively, the referents of any nouns or pronouns must be obvious, at least to the degree usually required in communication. In this text, however, the author employs the first person plural pronouns 'we' and 'us' in contrast to the noun 'robots'. As in: 'Robots are just like us' and 'we have rights'. Such pronouns refer to the author and the audience together, yet in this case, the author is actually one of the robots and not one of the group which has rights, i.e., humans. This makes the words 'us' and 'we' vague and possibly nonsensical, rendering the argument an example of a Fallacy of Clarity. If we were to consider the text more broadly, we might also call this a Fallacy of Consistency, since the author uses first person pronouns to refer to itself specifically as an AI in other places. Indeed, in the same paragraph we find the statement 'we don't want that' which is at best ambiguous and at worst a use of 'we' to mean AI robots in the sentence before 'we' is used to mean humans.

At this point we may wish to consider whether a reformulation of the argument will render it acceptable. A canonical realization without the pronoun problem would be:

> Robots should be given rights by people, because robots are just like people.

The analysis of Process and Reasoning would proceed in much the same fashion, and while a Language analysis might want to investigate further what exactly was meant by 'rights' here, the argument is in a general sense respectable and can be accepted as part of the debate on human/AI relations, even if it somewhat off-topic for this article.

4. Conclusion

The question posed in the title of this article is: 'How persuasive is AI-generated Argumentation?' In order to provide an answer, we now return to the two research questions asked in the introductory section. (RQ 1) What types of argument did GPT-3 produce? The analysis shows that the text generator was able to produce a variety of argument types, readily identifiable as established patterns of human reasoning. These included arguments from criterion, definition, similarity, equality, and authority: informally, we might describe these as arguments that something is true because something else is said of the same thing, because the same is said of something similar, or because someone says it is true. All of these are familiar forms of human argument, which suggests that GPT-3 is capable of employing the language of reasoning in natural ways.

The second question provides a sterner test: (RQ 2) Are these arguments acceptable? Of the five argument patterns examined, only one was considered to be an acceptable argument. The first was found to have committed a Productivity Fallacy: that is, it failed to produce new information in its conclusion and was guilty of what is commonly known as 'begging the question'. The second argument was of doubtful pertinence, but in any case, was judged to have committed a Lever Fallacy, since there was insufficient relevant likeness for an argument from similarity to have force.

The third argument was the only one considered to be acceptable. This argument from definition inferred soundly that a machine programmed by humans was a servant of humans. This is an example of an argument structure which is acceptable in itself, but at the same time very unlikely to be persuasive when it meets counter-arguments. A human arguer would perhaps not put forward an argument that rests on the assumption that what humans control cannot harm humans. In this case, it seems that even where GPT-3 does produce a cogent argument structure, its lack of broader knowledge and understanding, a common sense realization that humans can harm other humans, leads it to offer an argument which would face immediate rebuttal.

Argument four is another which commits a Lever Fallacy. The argument has the form of an argument from authority, but the cited authority is too vague and insubstantial to lend any force to the argument. The fifth and final argument contained some confusion over the identity of the arguer and its relation to the other entities under discussion. This led to a finding of Fallacy of Clarity, since it was not clear to whom the arguer was referring. It is probably fair to say that this form of mistake marks the author of the text as non-human more than any of the others: humans do usually have a grip on their own identity while arguing.

On top of the flaws already discussed, several of the arguments were suspected of failing the test of relevance by committing a Pertinence Fallacy. In spite of a clear set of instructions, the text generator interpreted the remit very widely and brought in arguments which, while related to the keywords of the task, did not address the actual issue at hand. It should be remembered too that a good deal of irrelevant writing was removed from the original eight outputs, suggesting that the overall proportion of genuinely relevant arguments produced was very low. This could be a sign of the problem that while a machine like GPT-3 can recognize and process language, it cannot understand it, and cannot tell if related material is related in a way which humans would consider relevant to the matter at hand. All of which means that the arguments generated are not likely to be persuasive.

The analysis has shown that GPT-3-generated arguments fall down at each of the three stages of argument evaluation. This suggests that the problems are not confined to one area, but are more pervasive. On the basis of our analysis, we would suggest three main areas which require improvement:

1. Relevance. Producing a text on a general topic is not sufficient when a particular position is supposed to be defended. Text generators must develop a better ‘understanding’ of what is required as part of the argument process.
2. Inference strength. GPT-3 proved quite adept at employing the language of argument forms. Information concerning what is necessary for the levers of those forms to possess sufficient strength is required in order to prevent it producing such weak reasoning.
3. Identity. All human arguers present their standpoints against a background of personal identity, beliefs, and experience; they are not only aware of who they are as individuals but also of their roles relative to others, in particular their audience. This is clearly expressed through their language and described in the study of rhetoric. A new rhetoric of machines may be required.

All three of these proposals can, we believe, be achieved in part through the integration of the CAPNA system into argument generation programming. The system would provide both a means of assessing arguments in text received by the machine and a filter which would prevent the generation of unacceptable or weak arguments. A machine which could identify the argument type present in text which it encountered or produced, and apply the relevant Procedural Questions, would have a strong indication of how humans would evaluate the text. This would mean giving greater credence to the information established by strong arguments in the learning process, which might also mitigate against radicalization effects.

Another important concern for us was the ability of the CAPNA and associated ATIP to perform in the analysis of a text of this kind. Here, we are fully satisfied that the procedure was adequate to the task and that the evaluations produced were robust and largely in line with our intuitive judgements. Although Argument 5 contained a flaw of a type not usually associated with human arguers, the PQs were still able to identify the weakness and reject it. We have found little to suggest that arguments produced by machine text generation are likely to differ from natural human language arguments in ways significant to their evaluation. It is worth noting once again, however, that the procedure described in this paper is designed only for the analysis of individual premise/conclusion pairs. Judging from the occasional contradictory statements found within the text, there is good reason to think that consistency of viewpoint throughout a longer discourse is another area where AI arguers will need to be improved. Evaluating such continuity is also a task for which the CAPNA requires further development.

Another area of research which could be explored is the use of comparative studies involving different systems of both identification and evaluation. The results of such studies would give a broader understanding of both the arguments produced by AI and of the tools we have developed. It would also provide more solid recommendations in cases where the identification and evaluation converge. Such studies would also look at outputs from other text generators, such as IBM's Debater, thus providing more comprehensive data about AI argument production. Comparative studies might also be carried out involving texts from human and automated text writers, with similarities and differences discussed and analysed.

Whilst we acknowledge that the current state of automated language processing has not reached the level of competence required for employing the CAPNA in the same way as a human analyst, it is our belief that we have shown the value which a familiarity with the PQs of an argument evaluation procedure could have as a means of filtering out poor argumentation. The challenge of using this information in the production of arguments will require further research and, crucially, close cooperation between argumentation theorists and computer scientists: understanding of what it would mean to have genuinely persuasive AI text generators loose in our society will also be aided by communication between technically and philosophically minded researchers.

Funding

This paper has been supported by an STSM grant from COST Action CA 17132 – European Network for Argumentation and Public Policy Analysis, and a grant from the Scientific Development Fund of the Faculty of Philology, University of Lodz.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

References

[1] W. Benzon, GPT-3: Waterloo or Rubicon? Here be Dragons, 2020, https://ssrn.com/abstract=3667608, Accessed 08 November 2021.

[2] G. Branwen, GPT-3 Creative Fiction, 2020, https://www.gwern.net/GPT-3, Accessed 08 November 2021.

[3] T.B. Brown, B. Mann, N. Ryder et al., *Language Models Are Few-Shot Learners*, 2020, arXiv:2005.14165v4.

[4] T.M. Conley, *Rhetoric in the European Tradition*, University of Chicago Press: London and, Chicago, 1990.

[5] J. Goodwin, Accounting for the appeal to the authority of experts, *Argumentation* **25** (2011), 285–296. doi:10.1007/s10503-011-9219-6.

[6] GPT-3, A robot wrote this entire article. Are you scared yet, human?, The Guardian 08 Sept. 2020, https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

[7] F. Grasso, A. Cawsey and R. Jones, Dialectical argumentation to solve conflicts in advice giving: A case study in the promotion of healthy nutrition, *International Journal of Human-Computer Studies.* **53**(6) (2000), 1077–1115. doi:10.1006/ijhc.2000.0429.

[8] N. Green, R. Dwight, K. Navoraphan and B. Stabler, Natural language generation of transparent arguments for lay audiences, *Argument and Computation* **2**(1) (2011), 23–50. doi:10.1080/19462166.2010.515037.

[9] Guardian US opinion editors, How to edit writing by a robot: A step-by-step guide, 11 Sept. 2020, https://www.theguardian.com/technology/commentisfree/2020/sep/11/artificial-intelligence-robot-writing-gpt-3.

[10] D.J. Gunkel, The other question: Can and should robots have rights?, *Ethics and Information Technology* **20** (2018), 87–99. doi:10.1007/s10676-017-9442-4.

[11] D.J. Gunkel, The right (s) question: Can and should robots have rights?, in: *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences*, B.P. Goecke and A.M. Rosenthal-von der Pütten, eds, Mentis Verlag, Paderborn, 2020, pp. 255–274. doi:10.30965/9783957437488_017.

[12] J. Harris and J.R. Anthis, The moral consideration of artificial entities: A literature review, *Science and Engineering Ethics* **27** (2021), 53. doi:10.1007/s11948-021-00331-8.

[13] M. Hinton, Overcoming disagreement through ordering: Building an epistemic hierarchy, *Studies in Logic, Grammar and Rhetoric.* **55**(1) (2018), 77–79. doi:10.2478/slgr-2018-0029.

[14] M. Hinton, Why the fence is the seat of reason when experts disagree, *Social Epistemology* **33**(2) (2019), 160–171. doi:10.1080/02691728.2019.1577512.

[15] M. Hinton, Towards a theory of informal argument semantics, in: *Reason to Dissent – Proceedings of the 3rd European Conference on Argumentation, Volume II*, C. Dutilh Novaes, H. Jansen, J. van Laar and B. Verheij, eds, College Publications, London, 2020, pp. 279–392.

[16] M. Hinton, *Evaluating the Language of Argument*, Springer, Cham, 2021.

[17] M. Hinton, An Informal Argument Pragmatics – Evaluating Argumentation Processes, *Journal of Prgamatics* (forthcoming).

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

[19] K. McGuffie and A. Newhouse, The Radicalization Risks of Gpt-3 and Advanced Neural Language Models, 2020, https://arxiv.org/pdf/2009.06807.pdf.

[20] A. Meleagrou-Hitchens and N. Kaderbhai, in: *Research Perspectives on Online Radicalisation: A Literature Review, 2006–2016. International Centre for the Study of Radicalisation*, Vox-Pol, Dublin, 2017.

[21] OpenAI, OpenAI API, 2021, https://openai.com/blog/openai-api/, Accessed 30 October 2021.

[22] A. Pelizzon and M. Gagliano, The sentience of plants: Animal rights and rights of nature intersecting, *Australian Animal Protection Law Journal* **11** (2015), 5–11.

[23] H.T. Tavani, Can social robots qualify for moral consideration? Reframing the question about robot rights, *Information.* **9**(4) (2018), 73. doi:10.3390/info9040073.

[24] F.H. van Eemeren, B.J. Garssen, E.C.W. Krabbe, A.F. Snoeck Henkemans, H.B. Verheij and J.H.M. Wagemans, *Handbook of Argumentation Theory*, Springer, Dordrecht, 2014.

[25] F.H. van Eemeren, R. Grootendorst, *A Systematic Theory of Argumentation*, Cambridge University Press, Cambridge, 2004.

[26] J.H.M. Wagemans, The assessment of argumentation from expert opinion, *Argumentation.* **25** (2011), 329–339. doi:10.1007/s10503-011-9225-8.

[27] J.H.M. Wagemans, Een systematische catalogus van argumenten [a systematic catalogue of arguments], *Tijdschrift voor Taalbeheersing* **36**(1) (2014), 11–30. doi:10.5117/TVT2014.1.WAGE.

[28] J.H.M. Wagemans, Constructing a periodic table of arguments, in: *Argumentation, Objectivity, and Bias: Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation*, P. Bondy and L. Benacquista, eds, OSSA, Windsor, 2016, pp. 1–12, https://scholar.uwindsor.ca/ossaarchive/OSSA11/papersandcommentaries/106/.

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

[30] J.H.M. Wagemans, Why missing premises can be missed: Evaluating arguments by determining their lever, in: *Proceedings of OSSA 12: Evidence, Persuasion & Diversity. Windsor, OSSA Conference Archive*, J. Cook, ed., 2020, https://scholar.uwindsor.ca/ossaarchive/OSSA12/Saturday/1.

[31] J.H.M. Wagemans, Argument Type Identification Procedure (ATIP) – Version 4, Published online December 30, 2021, www.periodic-table-of-arguments.org/argument-type-identification-procedure.

[32] D. Walton, *Appeal to Expert Opinion: Arguments from Authority*, Penn State Press, University Park, 2010.