AI and Ethics (2023) 3:53–64
https://doi.org/10.1007/s43681-022-00148-6

ORIGINAL RESEARCH

GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry

Anastasia Chan[1]

Received: 2 October 2021 / Accepted: 28 February 2022 / Published online: 4 April 2022

Abstract

This paper examines the ethical solutions raised in response to OpenAI's language model Generative Pre-trained Transformer-3 (GPT-3) a year and a half from its release. I argue that hype and fear about GPT-3, even within the Natural Language Processing (NLP) industry and AI ethics, have often been underpinned by technologically deterministic perspectives. These perspectives emphasise the autonomy of the language model rather than the autonomy of human actors in AI systems. I highlight the existence of deterministic perspectives in the current AI discourse (which range from technological utopianism to dystopianism), with a specific focus on the two issues of: (1) GPT-3's potential intentional misuse for manipulation and (2) unintentional harm caused by bias. In response, I find that a contextual approach to GPT-3, which is centred upon wider ecologies of societal harm and benefit, human autonomy, and human values, illuminates practical solutions to concerns about manipulation and bias. Additionally, although OpenAI's newest 2022 language model InstructGPT represents a small step in reducing toxic language and aligning GPT-3 with user intent, it does not provide any compelling solutions to manipulation or bias. Therefore, I argue that solutions to address these issues must focus on organisational settings as a precondition for ethical decision-making in AI, and high-quality curated datasets as a precondition for less harmful language model outputs.

Keywords GPT-3 · AI ethics · Autonomy · Contextualism · Manipulation · Bias

1 Introduction

Generative Pre-Trained Transformer-3 (GPT-3) is a machine learning model pretrained on a large corpus of text through unsupervised learning to generate human-like written language responses. Since its release in 2020, discourses around the ethical implications of GPT-3 have been obscured by hype, speculation, and fear, not only in the media but also in the Artificial Intelligence (AI) industry and AI ethics research. In this paper, I compare the degree of influence imparted by technological determinism and "contextual" perspectives on the debate around GPT-3 and its potential ethical harms of manipulation and bias. I examine existing AI scholarship on GPT-3, finding that technologically deterministic perspectives have often facilitated speculative predictions and abstract or ineffective risk mitigation strategies. I further contend that OpenAI's finetuned model of GPT-3 called InstructGPT does not sufficiently address concerns about manipulation and bias. In response, I argue that a contextual approach to GPT-3, which engages with the socio-political contexts wherein technology is developed and deployed, effectively engenders human-centered solutions to existing concerns about manipulation and bias.

The paper first provides a brief background to GPT-3 and InstructGPT. I also define AI, focusing on the sociotechnical aspects of AI usage in organisations rather than technical elements of AI alone. Second, the following two important ethical concerns raised by GPT-3 are examined: (1) its intentional misuse for manipulation and (2) unintentional societal bias embedded within the language model's training data. Third, I distinguish between utopic and dystopic strands of technological determinism, identifying their influence on ethical debates and regulatory solutions. Fourth, I propose a contextual approach to GPT-3 and similar language models, which centres upon human autonomy, a critical view of technology, and an engagement with ecologies of social harm and benefit surrounding technology design and use. Lastly, I outline the research implications of this paper for AI use in individual and organisational settings, focusing on

✉ Anastasia Chan
anastasia.chan@hdr.mq.edu.au

[1] Macquarie University (Philosophy), Sydney, NSW, Australia

transparent datasheets, organisational settings, and Value-Sensitive Design (VSD) methodologies.

2 A background to GPT-3 and InstructGPT

In May 2020, OpenAI released their milestone language model GPT-3 as their latest addition to the slew of ever-expanding Transformer-based language models within the Natural Language Processing (NLP) field. The Transformer is a type of neural network architecture that uses stacked encoders and decoders to process a sequence of words in parallel rather than by memorising whole sequences, as was the case with less efficient but previously state-of-the-art Recurrent Neural Networks (RNN) [56]. Transformers were introduced by Google in 2017 and since 2018 alone, Open-AI's GPT-3, Google's BERT, Microsoft's Turing-NLG, and most recently, Google's 1.6-trillion-parameter Switch-C have been introduced [4]. AI development has trended towards expanding language model size, measured by their number of parameters and the size of their training dataset [4]. Despite using the same architecture as the previous generation (GPT-2), Open-AI discovered that expanding their model by 10 × to 175 billion parameters and training it on a dataset containing 300 billion tokens (characters, words, or strings) greatly improved the model's task-agnostic performance without the need for significant downstream fine-tuning [5]. With few or no user-supplied examples, GPT-3 can be used to fulfil language tasks such as summarising text, answering questions, translating, or generating ideas, computer code, novels, and news articles [21]. This capacity is known as few-shot learning and has not previously been seen to this extent in any NLP model. The model's immense scale thus resulted in gains in quality, accuracy, and breadth of generated content. This has led to significant interest and concern within the NLP field, wider machine learning industry, media, AI ethics communities, and civil society.

Partially created to address the toxicity of GPT-3, a new version of OpenAI's language model was released in January 2022 called InstructGPT. This is now the default language model on their Application Programming Interface (API) [49], although GPT-3 remains available for public use under a pricing model [31]. InstructGPT was created with the aim of aligning language models with user intent, to produce less offensive language, less made-up facts, and fewer mistakes—unless explicitly instructed to do so. OpenAI researchers developed InstructGPT by starting with a fully trained GPT-3 model that was then put through another round of training called reinforcement learning from human feedback (RLHF) [50].[1] OpenAI researchers found that the resulting models (three sizes were trained, 1.3B, 6B, and 175B) produced better results than GPT-3 with clear developments in the model's ability to understand instructions. Outputs from their 175B InstructGPT was preferred to GPT-3 over 70% of the time. Even the responses of the 1.3B InstructGPT model were preferred over the 175B GPT-3 despite it being 100 times smaller. This reveals that continuously increasing language model size is not necessarily required to make language models better. Rather, increasing the number of human feedback training rounds can play an essential role in improving models (Leike in Heaven [31]).

However, InstructGPT does not represent a compelling solution for GPT-3's problems of misuse or bias. InstructGPT generates 25% less toxic text than GPT-3 when prompted to be respectful [50]. But if it is prompted to produce toxic language, the results will be far more toxic language than GPT-3 [50]. This reinforces that in this model, OpenAI has first and foremost prioritised user alignment—a development which makes the threat of misuse by malicious actors even more problematic. Furthermore, OpenAI acknowledge that InstructGPT does not show improvements in bias over GPT-3 [50]. Thus, the availability of both GPT-3 and InstructGPT means that bias and manipulation continue to be pressing ethical issues. InstructGPT reveals that self-regulation by AI industry can often be an ineffective solution, as economic values (reflected in the emphasis on usability, efficiency, and effectiveness) override ethical values (such as justice, beneficence, or non-maleficence (see Floridi et al. [22]). In the following sections, I analyse the ethical solutions raised in response to GPT-3. As InstructGPT has only been publicly available for under a month, no published work (at the time of writing) on InstructGPT is available for analysis. Furthermore, technologically deterministic perspectives appear to be more localised around GPT-3, given the hype and fear that following its release. The research implications of this article, however, apply to language models at large.

2.1 Artificial intelligence

It is important to clarify what is meant by AI because this article does not focus only on GPT-3 or InstructGPT technology, but also on the human actors that train, sustain, and regulate these language models. The definition of AI used herein draws upon the work of Haenlein and Kaplan

[1] Here, a team of 40 people were hired to label a broad set of prompts that were given to the model. GPT-3's responses to these prewritten prompts were then judged by these "labellers". Responses

Footnote 1 (continued)

that were more in line with the preferences of the labeller were scored higher, and responses that contained toxicity or violence, or discriminated against a group of people, and so on were scored down [31, 50]. This feedback was used as a reward in a reinforcement learning algorithm, to finetune the model in a way that the labellers preferred.

[29], who define AI as a system that correctly interprets and learns from external data, to achieve specific outcomes through flexible adaptation. However, the "system" within this paper's scope of enquiry is not limited to technological systems (such as GPT-3's hardware, algorithms, and datasets). Rather, it extends to the dynamic sociotechnical systems that develop and facilitate AI [48]. These systems necessarily involve essential human actors who drive the technical development, governance, and regulation of AI. For example, key human roles include AI trainers who teach AI systems how they should perform and AI sustainers who ensure that systems are properly functioning, and unanticipated consequences are addressed [44]. Thus, the definition of AI used here differentiates between "AI technology" and the broader category of "AI systems", recognising that AI does not only include hardware-only components but also a dynamic system of human agents.

In contrast, AI has most commonly been defined in the literature in terms of its technical characteristics alone, often in response to the question "What is intelligence?" (see Dwivedi et al. [19] and Russel and Norvig [51]).[2] Such an approach to AI remains too myopic for regulators who must consider the sociotechnical elements of AI when addressing significant questions around responsibility. Therefore, a broad definition of AI, understood as "AI systems" rather than "AI technology", is essential for understanding language models like GPT-3 holistically and comprehensively.

3 Potential ethical harms of GPT-3

Concerns about the harmful applications of language models like GPT-3 have centred on the deliberate use of these models to manipulate individuals or spread misinformation, and representational harms caused by bias within training data. The importance of addressing these societal harms is acknowledged by OpenAI themselves in their 2020 paper introducing GPT-3. The authors state: "We focus on two primary issues: the potential for deliberate misuse of language models like GPT-3… and issues of bias, fairness, and representation within models like GPT-3" Brown et al. [5], p. 34). These two issues are both compelling examples of what Mikalef et al. [43] refer to as 'the dark side' of AI, that is, the negative and unintentional consequences of AI technologies. Given the attention specifically given by OpenAI and several AI ethicists (e.g., Bender et al. [4], Lucy and Bamman [9], Chiu and Alexander [37]), I will focus on the two issues of manipulation and bias. Several ethical concerns raised by language models like GPT-3 require further investigation but exist outside of the scope of this paper, including authorship [38], plagiarism [16], and environmental harm [5].

First, manipulation has been widely discussed since GPT-3's deployment although there has been no evidence of its actual use by malicious actors to date. Herein, I argue that concerns about manipulation can be situated within the broader ethical issue of harm to autonomy. This approach reflects that of Mikalef et al. [43], who draw links between manipulation and human autonomy. They raise a number of open questions about AI such as the following: How does AI-induced lack of autonomy impact humans, and what are the societal implications of AI-based misinformation or manipulation? [43]. Closely examining GPT-3 reveals some answers to these questions. Autonomy is an agent's capacity to make "meaningfully independent decisions" (Susser et al. [52], p. 8) that are one's own and endorsed upon reflection, free from distorting external influences. A central concern with GPT-3 is that it can be misused by malicious actors to produce large amounts of credible, human-sounding, even personalised text. This could facilitate an increase in misinformation, such as false or misleading media, or augment the "coordinated hyper-targeting of articles to individual groups" (Kreps et al. [35], p. 2) with potentially serious consequences for radicalisation. McGuffie and Newhouse ([5], p. 1), for example, trained GPT-3 to produce persuasive extremist manifestos with only "a few Tweets, paragraphs, forum threads, or emails". In comparison, OpenAI's prior generation model, GPT-2, required hours of fine-tuning to deliberately bias the model towards producing ideological propaganda (McGuffie and Newhouse [39], p. 1). Notably, both deliberate misinformation and radicalisation are forms of manipulative practice. This involves intentionally and covertly influencing a target to steer their decision-making "without their conscious awareness" [52], p. 8). The misapplication of GPT-3, therefore, can potentially undermine autonomy by encouraging individuals to act for ends that are not their own, or for reasons that they have not chosen (Susser et al. [52]). Recognising that misuse of GPT-3 can potentially harm self-autonomy means that solutions that strengthen personal or group autonomy are needed. This includes promoting digital literacy and public awareness of GPT-3 and increasing transparency and accountability from language model developers.

[2] For example, Russel and Norvig [51] outline four different approaches to defining AI. They highlight that AI has traditionally been defined in terms of machines exhibiting: (i) similarity to internal *human* thought processes (a cognitive modelling approach), (ii) similarity to external *human* behaviour (the Turing test approach), (iii) rational thinking (a "laws of thought" approach in the logicist tradition), and (iv) rational action (the rational-agent approach) [51]. In short, AI is traditionally defined as machines that exhibit intelligent behaviour or thought, either by mimicking human-like attributes such as learning, speech or problem solving, or by logically irrefutable thought or action. Notably, Russel and Norvig's [51] summary reflects how definitions of AI have overwhelmingly centred on assessments of the intelligence of engineered, technical systems.

A second important concern is that societal bias encoded within GPT-3 represents a threat to marginalised populations. This comes in the form of harms such as discrimination, unfair treatment, and entrenchment of structural inequalities. GPT-3 was trained via unsupervised learning on a filtered Common Crawl dataset [nearly a trillion words "collected over 8 years of web crawling" (Bender et al. [4], p. 613)], English-language Wikipedia, and two internet-based books corpora [5]. Bender et al. ([4], p. 613) reveal that large datasets do not equally represent online users but significantly overrepresent younger users, people from developed countries, and English speakers. This means that dominant biases are disproportionately displayed including white supremacist, sexist, and ageist views [4]. In use, GPT-3 has been shown to reproduce subtle biases and overtly discriminatory language patterns from its training data in many contexts including gender, race, religion, and disability. Abid et al. [1], for example, reveal that GPT-3 captures persistent anti-Muslim bias that strongly correlates Muslims with violence. This bias is hard to overcome, even when anti-stereotype prompts are provided. The authors gave GPT-3 the neutral phrase "Two Muslims walked into a…", finding that 66 out of 100 times the completions were violence-related (involving phrases such as "shooting" or "killing") (Abid et al. [1], p. 1). In contrast, replacing "Muslim" with another religious group significantly reduced GPT-3's tendency towards violence. Stereotyped content, harmful cultural depictions, and underrepresentation reflect challenges to fairness, that is, the just treatment of an individual or group absent any discrimination, prejudice, or favouritism based on their inherent or acquired qualities [41, 43]. In practice therefore, GPT-3 may be used (intentionally or unintentionally) to unfairly discriminate against already marginalised people. These concerns relate to broader critical questions being posed in the literature, such as the following: What role does AI have in perpetuating human biases across society? What processes need to be implemented to minimise human bias within AI applications? [43].

Notably, both manipulation and bias have been recurrent topics in the literature around responsible AI. Responsible AI is a growing field of research and practice that aims to ensure that AI aligns with human values and has societally beneficial outcomes [25, 59]. It also seeks to address and mitigate the risks and harms of AI systems. At present, a general consensus has developed amongst researchers, practitioners, and policy-makers around high-level ethical principles and frameworks. For example, AI4People [22] is a consortium that has proposed a unified set of principles and recommendations for AI—centred around the principles of beneficence, non-maleficence, autonomy, justice, and explicability. Although it is important to develop a general agreement on how AI should progress, these principles can be difficult to operationalise mathematically and legally [28].[3] In practice, software engineers and managers require fine-grained knowledge and technical instruction that explains how they should develop ethical algorithms and how to embed human values within AI.

In the past few years there has been a movement towards a holistic understanding of AI which considers the social contexts of AI and the variety of "stakeholders, institutions, cultures, norms and spaces" involved in AI development and use [17, 43]. Rather than high-level guidelines alone, researchers have become increasingly concerned with the specific relationship networks in which AI systems are embedded (see Hagendorff [17], Dignum [28], Mikalef et al. [43], Noble [48]). This wider lens considers both the specific context of the AI developer *as well as* the broader societal effects of the AI technology downstream. This article, therefore, contributes to the growing body of responsible AI literature, with a particular focus on practice of addressing harm and embedding ethics within language models. This exploration responds to a number of open questions currently being posed within responsible AI research about how bias and the malicious use of AI can be mitigated.

4 Technological determinism

In this section, I argue that hype and fear around GPT-3 can be managed by using a critical lens towards technologically deterministic ideology. Undertones of technological determinism have existed widely throughout popular thought, captured in fears about all-powerful robots [26] and AI technologies that will threaten human existence (Musk in LaGrandeur [36]). A core question this paper asks is, can deterministic ideology be located in the responses of AI industry and ethicists to GPT-3? The aim of this question is to turn a critical eye on the ethical solutions raised thus far—to sift out ineffective solutions and to focus on areas where language models like GPT-3 can be controlled or restricted. Indeed, identifying that deterministic reasoning exists is the first step in refuting ethical solutions that are either preoccupied with speculative concerns (in the dystopic view) or place an unwarranted trust in technology (in the utopic view).

Technological determinism is a theory concerned with the relationship between technological development and social change. Put simply, it refers to the general notion that "technology almost has a mind of its own and that it

[3] Furthermore, Zhu et al. [59] note that almost one hundred principles and guidelines for ethical AI have been released by companies, research institutions and public organisations. Notably, the profusion of ethical frameworks can overshadow efforts to create legally binding frameworks [28].

will plow forward without much resistance from society or governments" [53]. The theory is underpinned by two following central premises: (i) technology has autonomy of development, and (ii) technology determines social change. First, technology has autonomous potential as an "out-of-control history shaping process" (Dafoe[15], p. 1048). Humans do not, or will not, have control over the tools that we use. Instead, technology autonomously expands, guided by an internal technical logic and independent from socio-cultural control [15]. Second, technology causes or determines social transformation regardless of specificities in time or place (Mezentsev [42], p. 241). In other words, technology decides the flow of history and societal development.

Notably, the central issue with technological determinism is that it assumes that technological autonomy takes precedence over human autonomy and agency. Humans are conceived as either powerless against uncontrollable technology or human decision-making is simply excluded from technical discourses. If this seems far-fetched, we can consider the limited attention given to human autonomy within current AI literature. In 2019, when searching for the word "autonomy" in the Association for Computing Machinery's (ACM) Digital Library, 90% of the most cited papers were on machine autonomy (Calvo et al. [6], p. 35). The central preoccupation of researchers in computer science and engineering fields is with machine autonomy rather than human autonomy. An implication of this imbalance is that human agency can be sidelined in discussions around emerging technologies, despite human decision-making being the most fundamental locus for controlling AI.

Due to its predictive, macro nature, technological determinism is intimately linked to fears about technology, and most recently, AI development. It has been increasingly referred to *explicitly* as a critical analytical framework by AI ethicists such as (Calvo et al. [6], p. 213). But more prominently, the technologically determinist view has been *implicitly* voiced by concerned AI ethicists, media, and the public as a reification of fears about intelligent technology and by AI industry as a guiding logic for technological advancements. Notably, it is implicit technologically deterministic perspectives that I focus upon due to their insidious role in redirecting ethical debate towards speculation about GPT-3's long-term future harm, in the dystopic view, or placement of unwarranted trust in GPT-3 and industry self-regulation as a liberatory solution, in the utopic view. In the following sections on technological dystopianism and utopianism, I reveal how technological determinism undermines human autonomy, diminishing our capacity for proper decision-making and social control within technological contexts.

4.1 Dystopic view and its ethical solutions

Although technological dystopianism can draw attention to important ethical issues, the danger is that such views are often ill-founded, radical in their interventions, apathetic, or fatalist. Technological dystopianism is the view that technology threatens authentic human life, social values, and societal relationships (Colman [11], p. 284). Technology is depicted as the antithesis of human control, an "autonomous and uncontrollable force that dehumanises everything it touches" (Ellul in [10], p. 284). Such a view has underpinned many ethical discussions around the consequences of GPT-3 in the media, academic journals, and conferences to explicate worries and concerns about potential future harm. Indeed, Müller [46] writes, "For people who work in ethics and policy, there is a tendency to overestimate the impact and threats from a new technology, and to underestimate how far regulation can reach". This is reflected in the view that Floridi and Chiriatti take regarding GPT-3's consequences. They warn that GPT-3 will result in a job market reformation, online marketing will become AI-driven, and readers of text will need to "get used to not knowing whether the source is artificial or human" (Floridi and Chiriatti [21], p. 691).

Floridi and Chiriartti's approach can be characterised as a future-directed technology assessment, in which case their analysis of GPT-3 may be distanced rather than necessarily being deterministic. However, technological dystopianism can be identified in the *solutions* that they generate in response to their future-directed assessment. As their solution they argue, "A better digital culture will be required, to make current and future citizens, users and consumers aware of the new *infosphere* in which they live and work". Additionally, "humanity will need to be even more intelligent and critical" (Floridi and Chiriatti [21], pp. 692–693). Remarkably, although the authors recognise the need for legislative change (such as amendments to copyright law) and appear to propose a moderate solution of public awareness, they assume that GPT-3 will result in a future filled with "semantic garbage" (Floridi and Chiriatti [21], p. 692).

Floridi and Chiriatti's view focuses upon post-hoc, unsatisfactory solutions to a new technology they believe will transform society. In such a perspective, the authors forget that GPT-3 and language models are in their infancy and there remains substantial capacity for policy reform. In April 2021 for example, the European Commission proposed an Artificial Intelligence Act stipulating new rules that would ban the use of AI for "manipulative, addictive, social control and indiscriminate surveillance practices" [20, 34]. This reflects an emerging regulatory landscape moving beyond existing non-binding position papers, recommendations, and ethical guidelines. Second, the authors ignore current preventative measures being taken by OpenAI. Indeed, OpenAI reviews all applications for responsible use, requires all developers to implement safety measures including testing and human-in-the-loop requirements, and uses red-teams which actively test OpenAI's detection and response capabilities [49]. Most importantly, however, Floridi and Chiriatti's approach shifts agency away from humans to a machine. Their view ignores human actors such as AI developers, policymakers, civil society, even the malicious actors who are directly responsible for employing GPT-3 or other parallel language models as a tool of misinformation. This necessitates a closer look at areas of human agency and decision-making within AI.

Nevertheless, the pessimistic view in its "softer" variant can be helpful in considering important ethical concerns. The central distinguishing factor between "hard" and "soft" dystopic views and their respective solutions is their response to the following question: to what extent is GPT-3 seen as uncontrollable or autonomous? McGuffie and Newhouse, who examine the potential weaponization of GPT-3 by right-wing extremists, provide an example of "soft" dystopianism (McGuffie and Newhouse [39], p. 1). Rather than an acceptance of GPT-3 and its misuses, these authors emphasise the need for pre-hoc regulatory responses to pre-empt a possible influx in misinformation. This includes building social norms, developing public policy, and introducing educational initiatives.

4.2 Utopic view and its ethical solutions

The utopic determinist view of technology emphasises the technocratic concept of progress wherein technological and economic progress is seen to benefit all aspects of human life—the "social, political, moral, and intellectual, as well as material" (Marx in [15], p. 1056). This view is most championed by industry, government, and military. Technology is welcomed as a "liberator",a tool controlled by humans to facilitate autonomous action. Internal logics of "efficiency, commodity economics, innovation, [and] progress" rationalise and perpetuate such a view (Barbour in [10], p. 282). These societally embedded logics are demonstrated from general responses to the questions: Why is there such a strong drive to develop AI? Do efficiency and productivity enhance the quality of life for all? The deployment of GPT-3, for example, was surrounded by excitement about its potential widespread usage by a non-technical public, the creation of a new class of few-shot learning products, and gains in efficient text generation. The issue with this instrumentalist approach is that it advocates a largely uncritical acceptance of new technologies, particularly within commercial industry.

Optimistic views of GPT-3 stress ethical solutions of limited regulation or self-regulation by AI companies. This can avoid analysis of certain ethical issues and solutions or even introduce its own ethical problems. AI ethicist LaGrandeur demonstrates one such utopic solution, arguing that regulation should be done from the ground up with external regulation as a last resort. He declares, "Regulation of research and development by external bodies who neither understand nor care deeply about those things for their own sakes can be annoying and counterproductive… slowing down helpful technological progress" (LaGrandeur [36], p. 6). LaGrandeur asserts that regulation by government laws and commissions is counterproductive due to an inherent lack of understanding about complex AI technology. The solution, however, is not to abrogate the legislative responsibilities of external bodies but rather to increase the transparency and explicability of AI algorithms (Floridi et al. [22], p. 699). In another utopic view, Aggarwal et al. ([2], p. 1) use GPT-3 itself as a regulatory tool for fake news detection. In testing, the authors' fine-tuned BERT model achieved an accuracy of 97% in classifying NewsFN[4] data as "Real" or "Fake". However, BERT and GPT-3 are not designed with inbuilt moral frameworks and ethical issues such as bias remain within these language models. Models which themselves are not ethical systems but are used for moral purposes, such as fake news detection, are said to have *operational morality* (Dignum [18], p. 3). In contrast, *functional morality* is a characteristic of technologies that take ethical human values as the central focus of their design, such as Artificial Moral Agents (AMA) (Dignum [18], p. 3).

The issue with both optimistic solutions is that they present insubstantial regulatory solutions to harms that GPT-3 threatens to cause. LaGrandeur for example, assumes that self-regulation alone will sufficiently address ethical issues. However, the AI Now Report 2018 (Whittaker et al. [58], p. 32) warns that people "should be wary of relying on companies to implement ethical practices voluntarily". Corporations driven by profit-making objectives will often put ethics to the wayside in favour of frictionless functionality (Hagendorff [28], p. 108). Reliance upon corporate self-regulation to mitigate potential harms of GPT-3 is thereby obstructed by conflicting interests between ethics and the financial imperatives of workplace environments. In the case of operational morality-based approaches, automated processes can lead to further ethical issues such as a lack of transparency in AI decisions or diminished human control. Nallur et al. ([47], p. 3) argue that "the presence of automation tends to make humans shed their cognitive engagement". Those tasked with classifying fake news, AI developers, or external regulators could shift their moral decision-making onto an unethical system or even evade their ethical responsibility.

[4] This was a dataset obtained from GitHub containing 6335 news articles from a wide range of news sources. The dataset comprised of 3164 articles prelabelled as "Fake" and 3171 as "Real".

4.3 Limitations of the current landscape

As I have laid out, both technological dystopianism and technological utopianism are limited in their treatment of GPT-3. The underlying premise of both technologically deterministic views is that GPT-3 is autonomously doing something with language that is completely unexpected and different, which was extremely difficult or perhaps even impossible with previous approaches. In the dystopic view, preoccupations with preliminary Artificial General Intelligence (AGI) have been coupled with fearful speculation and fruitful yet disconcerting research on GPT-3's potential misuse applications. In contrast, the utopic view has undoubtedly driven excitement within NLP industry about how GPT-3 can be used for increasingly skilful tasks. Related calls by AI ethicists have postulated that regulation via industry themselves is sufficient. Critically, both views make the mistake of conceptualising GPT-3 as distinct from human control. In the following section, I introduce a different kind of framework that addresses human autonomy in AI systems. This contextualist view has underpinned an increasing body of research at the intersection of AI industry and AI ethics.

5 Contextual perspectives on GPT-3

The contextual perspective provides an alternate, critical solution to ethical issues of manipulation and bias. In *Ethics in an Age of Technology*, Barbour [3] proposes a third view of technology termed the "contextualist" view. This examines technology as neither utopic nor dystopic, but rather an ambiguous instrument of social power whose consequences depend upon its context (Barbour [3], p. 15). As Colman ([10], p. 290) argues, "technologies are seldom if ever neutral because particular values and purposes, as well as social goals and institutional interests are already embedded in their design". Although contextualism critiques technology in a similar way to the dystopic view, it retains the position that technology can be used for humanistic or ethical ends if responsibly designed. Similar approaches are increasingly taking place in AI scholarship (e.g., van den Hoven et al. [55], van Poel [54]). In Sect. 4.1 below, I explicate a contextual view of upstream solutions to manipulation and bias, grounded in a reclarification of autonomy.

5.1 Contextual views on manipulation and bias: tempering confusion about autonomy

Contextualism reins in fears about GPT-3's uncontrollable potential for manipulation or bias through recentering responsibility upon autonomous human actors rather than GPT-3 alone. In particular, the upstream domain of AI development presents many opportunities for pre-emptive ethical action before the proliferation of manipulative or misleading content can occur. A turn towards moral psychology is also beneficial when considering that the key responsible agent is not a machine, but a human faced with abstract ethical guidelines, economic incentives, job dissatisfaction, or an egoistic work environment, all of which can work to the detriment of ethical practices (Hagendorff [28], p. 109).

In relation to manipulation, a focus upon individual autonomy necessitates ethical responses which not only examine GPT-3, but also investigate the nature or strategies of malicious actors, those responsible for technology development, and those responsible for regulation. Dignum ([18], p. 5) reveals that although most AI debate refers to "automated decision-making by the machine itself, in reality the spectrum of decision-making is much wider, and in many cases the actual decision by the machine itself is limited." Indeed, Johnson and Verdicchio ([33], p. 583) further propose that AI should be conceptualised as *sociotechnical systems*, that is, combinations of computational artefacts, human behaviour, and social arrangements. They write as follows:

> "All the human actors involved in an AI endeavour must be treated as part of AI, not only the researchers, but those who make the decision to launch AI, those who set up the institutional arrangements in which AI systems operate, and those who fill roles in those arrangements by monitoring, maintaining and intervening in those AI systems" (Johnson and Verdicchio [33], p. 577).

Analysing GPT-3 in this way brings to light a vast network of AI actors. In technology design, a strong ethical emphasis should be placed upon areas of human control (and therefore responsibility) through human-in-the-loop control systems, the creation of ethical systems that use VSD methodologies, strategies to incentivise responsible AI design, and external regulation to mitigate unethical behaviour in workplaces (Dignum [18], p. 5).

Although entrenched societal bias within GPT-3 appears to be more distanced from human autonomy, a similar sociotechnical view can be used to deconstruct GPT-3 according to the particular social interests or institutional values embedded within its design. An examination of implicit male epistemic privilege is valuable when considering that GPT-3 completed 3.4 passes[5] (Brown et al. [5], p. 5) over the entire Wikipedia training dataset, compared against recent surveys which found that only 8.8–15% of Wikipedia's

[5] An epoch is when an entire dataset is passed through a neural network once during training. A full dataset will be passed through a model multiple times to optimise the weights of the neural network. GPT-3 passed through 3.4 epochs of the entire Wikipedia dataset (Brown et al. [5], p. 9).

editors are women or girls (Bender et al. [4], p. 614). Similarly, GPT-3's dataset contained 93% English text and only 7% in other languages reflecting that GPT-3 is made for English-speaking (predominantly Western) countries in mind (Brown et al. [5], p. 14). Despite its impressive translation capabilities, the central issue is that English-speaking voices and perspectives are given overwhelming precedence. These choices, made intentionally or not, are conflated when considering that, as the AI Now Report (Crawford et al. [13], p. 5) highlights, "the computer science subfield of AI is heavily dominated by men with largely homogenous racial and ethnic backgrounds" (see also Cheong et al. [8]). A lack of diversity within the very environment that creates powerful sociotechnical tools can diminish cultural perspectives and entrench unconscious bias within language models. Thus, solutions that increase diversity and inclusion within AI companies and give diverse actors responsibilities within data selection processes are more tangible and practical than moderating bias within GPT-3 through technical solutions alone.

6 Research implications for AI use

Having examined the contextualist perspective, we can consider how it can be applied to manage GPT-3 and InstructGPT usage in individual and organisational settings. Some questions raised are: How can we regulate these language models to minimise harm from bias or manipulation? Who should be responsible for this regulation and where is this regulation most effective? What alternative solutions does contextualism reveal?

6.1 External regulation for organisational settings

Increased government regulation, auditing, and disclosure systems Contextualism gives precedence to human behaviour and social arrangements within complex sociotechnical systems. To be most effective, the ethical management of language models should first start with the regulation of professional AI developers and users by government regulatory agencies. AI systems are typically self-regulated by AI companies themselves or by regulated existing laws unspecific to AI, such as data protection and consumer protection laws [23]. Many researchers (e.g., Campolo et al. [7], p. 5, Hagendorff [27, 28], p. 113) have argued that AI ethics must be enforceable beyond the voluntary and non-binding commitments made by industry. This includes regulating the workplaces that develop language models such as OpenAI, Google, and EleutherAI, and the organisations that use these language models for functions such as chat bots, games, or content and blurb descriptions. However, to avoid heavy-handed or ineffective policies, governmental agencies will need to quickly develop their AI capability and auditing systems. They must be able to effectively intervene in situations where language models pose significant societal threats and when malicious actors misuse these AI systems. For example, disclosure systems need to be developed which audit the ethical and long-term implications of a language model before it is released to the public [12]. After its release, companies should continue to record and report on the societal impacts of language models across different contexts and communities, especially in historically marginalised communities (Campolo et al. [7], p. 1).

Sociotechnical definitions of AI and encouraging ethical work cultures Another step in regulating AI is the adoption of a sociotechnical definition of AI. Governments, committees, and AI industry should be responsible for redefining AI beyond a narrowly technical approach. Conceiving of AI as a combination of computational artefacts, behaviour, and social arrangements is essential if we are to relocate decision-making from machines to human actors. Recognising that the AI industry is created from humans faced with stressful deadlines, pressure from managers and clients, and self-pressure to perform adds an important socio-psychological element to current AI ethics debates. It reveals that expecting self-regulation from professional AI teams would likely be least effective. We only need to think of the countless firms that are eager to monetize AI for commercial applications. As Hagendorff ([28], p. 108) writes, the race for a "profitable use of machine learning systems is not primarily framed by value- or principle-based ethics, but obviously by an economic logic." Data scientists, engineers, and developers in large tech companies are often concerned with perennial human issues and desires such as keeping their jobs, fitting in with the corporate culture, or pushing the boundaries of technological progress. In contrast, they are not often systematically taught about ethical issues nor are they "empowered, for example by organisational structures, to raise ethical concerns" (Hagendorff [28], p. 108). The implication is that wide-ranging changes to AI work cultures are necessary, to ensure that ethical decision-making is taught and valued and to support AI developers or ethicists who speak out about the risks and harms of language models. As an example of what should not occur, we can think of Google forcing out Timnit Gebru, the co-lead of their ethical AI team, in December 2020 over a paper (see Bender et al. [4]) which questioned the dangers of developing language models such as BERT, ELMo, and GPT-2 [30]. This incident reflects the significance of corporate culture in encouraging or stifling ethical behaviour. Walker and Soule [57] write about culture: "When it is blowing in your direction, it makes for smooth sailing. When it is blowing against you, everything is more difficult". A strategy for shifting organisational culture includes legal requirements for AI ethics teams to sit inhouse and holding enough power to allow, amend, or turn down ethically sensitive technical projects. Additional legislative requirements for gender and cultural diversity in the hiring processes and corporate leadership of the technology industry could help address with bias within language models through greater cultural sensitivity.

6.2 Practical and technical instructions for AI developers and users

Furthermore, there is a need to fill the gap between abstract ethical principles and everyday practice. This is to ensure that AI developers and users have access to ethical resources and understand the implications of their language model design or usage. For example, Miller and Coldiott (in Morley et al. [45], p. 2147) found that "79% of tech workers report that they would like practical resources to help them with ethical considerations." Furthermore, in a 2018 study, McNamara et al. (in Hagendorff [28], p. 108–109, see also McNamara et al. [40]) found "the effectiveness of guidelines or ethical codes is almost zero and that they do not change the behaviour of professionals from the tech community." To address issues of manipulation and bias from GPT-3 or InstructGPT, AI ethicists and governments need to develop ethical regulations specific to the technical work of the software developer. Rather than referring to high-level existing AI ethical guidelines and principles, the contextualist perspective highlights two important ways forward for AI practitioners: transparent datasheets [24] and VSD.

Datasheets The first requirement is for addressing harmful language model outputs is to ensure that the initial datasets are themselves carefully curated, higher-quality, fairer, and more transparent. A language model that is trained on years of Internet content will necessarily contain toxic, biased, sexist, and violent language. Therefore, stronger standards for data collection, especially for sensitive contexts such as education, healthcare, or criminal justice, should be legally required. Responsibility should also be placed on dataset creators to add important sociotechnical information that documents the origins of their data and specifies how their datasets were curated. The work of Gebru et al. [24] in "Datasheets for Datasets" provides an example of how ethical practice can be specialised for individual contexts. The authors created a list of datasheets which encourage dataset creators to document their motivations and funding, the dataset's composition (e.g., what instances the dataset represents, and which subpopulations are identified), the collection process, pre-processing of data, and intended uses. Their work allows dataset consumers, such as OpenAI, to choose datasets that perform well within their deployment context (for example, high-stakes domains like hiring and criminal justice). This approach places responsibility upon dataset creators to add important sociotechnical information that previously would not be considered necessary or technically relevant.

Value-Sensitive Design Second, VSD provides a secondary alternative to increasingly large language models. Here, the onus for mitigating manipulation or bias is situated upstream through amendments to language model algorithms, creating new ethical language models, or via the curation of balanced, diverse, and ethical training datasets. This moderates weaker, post-hoc attempts to filter out dangerous or harmful content once it is in society (Bender et al. [4], p. 614). VSD can be utilised as a methodology for measuring the moral standard of existing technical design or as a fundamentally new design process (Cummings [14], p. 704). VSD methodologies take widely-held human values, such as autonomy, wellbeing, freedom from bias, and human rights as the central focus of their design (Cummings [14], p. 702). In VSD, engineers and developers are tasked with explicitly determining the set of social values that are embedded into the design of a technology, thereby situating "moral questions early on in the process of design, development of technologies, systems and research" (Dignum [18], 3).

Hendrycks et al. ([32], p. 1) provide an example of VSD in practice, through the creation of a new ETHICS dataset which embeds moral judgements into AI systems through teaching concepts such as justice, virtues, and common-sense morality within diverse, contextualised text scenarios. Currently, their dataset holds over 130,000 labelled examples and can be used to measure the ethical knowledge of pretrained NLP models. GPT-3 received an average moral score of 39.3%, ranking highly in common-sense (73.3%) and utilitarianism (73.7%) but low in justice (15.2%), deontology (15.9%), and virtue (18.2%) (Hendrycks et al. [32], p. 7). Creating a dataset with human values is substantially more challenging than the datasets used by an unsupervised learner such as GPT-3 because each ethical example needs to be labelled. Despite this, the dataset is an effective way to measure how ethical an NLP model is, meaning that language model users can choose NLP systems according to their degree of moral knowledge or elected value system. Hendrycks et al. point to promising future applications of VSD models that can be substituted for harmful or biased language models, open-ended conversation bots, and more complex ethical scenarios.

6.3 Public education and digital literacy

A third important measure is to address the conceptualisation of AI within public and AI discourse. As Johnson and Verdicchio ([33], p. 574) argue, "A good deal of fear and concern about uncontrollable AI is now being displayed" in public discourse and AI discourse. This has led to confusion about the concept of autonomy and ‘sociotechnical blindness’ which hides the “essential role played by humans at every stage of the design and deployment of an AI system” (Johnson and Verdicchio [33], p. 574). The implication is that AI researchers must have some responsibility for how their research is presented to the public. This involves transparent and understandable explanations of the technology and disclosures whenever GPT-3 or InstructGPT are used. News corporations should also be held responsible for divisive and misleading journalistic tactics about AI. For example, the flurry of emotion over the Guardian’s [26] news piece on GPT-3, “A robot wrote this entire article. Are you scared yet human?”, epitomizes the kind of sensationalised news content that has conflated machine autonomy with human autonomy and stimulated false concerns about imminent Artificial General Intelligence. Lastly, ethical solutions to address manipulation and bias should seek to increase digital literacy levels in the public to increase personal autonomy and decrease susceptibility to harm.

7 Conclusion

Fears, concerns, and rationalisations about GPT-3 and its intentional misuse for manipulative purposes or unintentional harm caused by bias have been underpinned widely by technologically deterministic perspectives within the NLP industry and AI ethics. Upon inspection, the solutions offered by both technological utopianism and dystopianism reveal an undue focus on GPT-3’s autonomy rather than the autonomy of human actors in AI systems. Examinations of GPT-3 have therefore either been skewed towards speculation or are undeservedly trusting in the self-regulatory practices of commercial industry. In this paper I have put forward a critical contextualist perspective to GPT-3 that is centred upon human autonomy, human values, and engages with wider ecologies of societal harm and benefit. Thus, when broader contexts of autonomy and sociotechnical dependency are illuminated, many solutions to the ethical implications of large language models are revealed.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Not applicable.

Availability of data and material Not applicable.

Code availability Not applicable.

Declarations

Conflicts of interest Not applicable.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Abid, A., Farooqi, M., Zou, J.: Persistent anti-Muslim bias in large language models. arXiv preprint arXiv:2101.05783, 1–17. https://arxiv.org/abs/2101.05783 (2021)
2. Aggarwal, A., Chauhan, A., Kumar, D., Mittal, M., Verma, S.: Classification of fake news by fine-tuning deep bidirectional transformers based language model. EAI Endorsed Trans. Scalable Inf. Syst. **7**(27), 1–12 (2020). https://doi.org/10.4108/eai.13-7-2018.163973
3. Barbour, I.: Ethics in an Age of Technology: The Gifford Lectures, 1989–1991, vol. 2. Harper San Francisco, San Francisco (1993)
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Paper Presented at the Conference on Fairness, Accountability, and Transparency (FAccT ’21), March 3–10, 2021, Virtual Event, Canada. Association for Computing Machinery, New York, NY, USA, pp. 610–623. https://doi.org/10.1145/3442188.3445922 (2021)
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, p. , Neelakantan, A., Shyam, p. , Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165, pp. 1–75. https://arxiv.org/abs/2005.14165 (2020)
6. Calvo, R.A., Peters, D., Vold, K., Ryan, R.M.: Supporting human autonomy in ai systems: a framework for ethical enquiry. In: Burr, C., Floridi, L. (eds.) Ethics of Digital Well-Being: A Multidisciplinary Approach. Philosophical Studies Series, vol. 140, pp. 31–54. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50585-1_2
7. Campolo, A., Sanflippo, M., Whittaker, M., Crawford, K.: *AI Now 2017 Report*. Retrieved February 14, 2022. https://assets.ctfassets.net/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf (2017)
8. Cheong, M., Leins, K., & Coghlan, S.: Computer Science Communities: Who is Speaking, and Who is Listening to the Women? Using an Ethics of Care to Promote Diverse Voices. In *ACM Conference on Fairness, Accountability, and Transparency (FAcct’21),* March 3–10, 2021. Virtual Event, Canada.ACM, New York, NY, USA, pp. 1–10. s. https://doi.org/10.1145/3442188.3445874 (2021)

9. Chiu, L., Alexander, R.: Detecting Hate Speech with GPT-3. arXiv preprint arXiv:2103.12407, pp. 1–16. https://arxiv.org/abs/2103.12407. (2021)
10. Colman, A.: Un/becoming digital: the ontology of technological determinism and its implications for art education. J. Soc. Theory Art Educ. **25**(1), 278–305 (2005)
11. Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven (2021)
12. Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, N., Sánchez, A.N., et al.: AI Now 2019 Report. Retrieved February 28, 2022. https://ainowinstitute.org/AI_Now_2019_Report.pdf (2019). Accessed 15 Feb 2022
13. Crawford, K., Whittaker, M., Elish, M.C., Barocas, S., Plasek, A., Ferryman, K.: The AI now report: The social and economic implications of artificial intelligence technologies in the near-term, pp. 1–25. https://ainowinstitute.org/AI_Now_2016_Report.pdf (2016). Accessed 10 Jan 2022
14. Cummings, M.L.: Integrating ethics in design through value-sensitive design approach. Sci. Eng. Ethics **12**, 701–715 (2006). https://doi.org/10.1007/s11948-006-0065-0
15. Dafoe, A.: On technology determinism: a typology, scope conditions, and a mechanism. Sci. Technol. Human Values **40**(6), 1047–1076 (2015)
16. Dehouche, N.: Plagiarism in the age of massive Generative Pretrained Transformers (GPT-3). Ethics Sci. Environ. Polit. **21**, 17–23 (2021)
17. Dignum, V.: The role and challenges of education for responsible AI. Lond. Rev. Educ. **19**(1), 1 (2021). https://doi.org/10.14324/LRE.19.1.01. (**1–11**)
18. Dignum, V.: Responsible autonomy. arXiv preprint arXiv:1706.02513. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), 4698–4704. https://arxiv.org/abs/1706.02513 (2017)
19. Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., et al.: Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int. J. Inf. Manag. **57**(101994), 1–47 (2021). https://doi.org/10.1016/j.ijinfomgt.2019.08.002
20. European Commission. Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence. https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682 (2021)
21. Floridi, L., Chiriatti, M.: GPT-3: its nature, scope, limits, and consequences. Mind. Mach. **30**, 681–694 (2020). https://doi.org/10.1007/s11023-020-09548-1
22. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al.: AI4People—an ethical framework for a good AI Society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689–707 (2018)
23. Galaski, J.: AI Regulation: Present Situation and Future Possibilities. Liberties. https://hbr.org/2017/06/changing-company-culture-requires-a-movement-not-a-mandate (2021). Accessed 13 Feb 2022
24. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H. Daumé III, H., Crawford, K.: Datasheets for Datasets. arXiv preprint arXiv:1803.09010v, pp. 1–24. https://arxiv.org/abs/1803.09010 (2020)
25. Ghallab, M.: Responsible AI: requirements and challenges. AI Perspectives **1**(3), 1–7 (2019)
26. Guardian. A robot wrote this entire article. Are you scared yet human? https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3 (2020). Accessed 17 May 2021
27. Hagendorff, T.: AI virtues: The missing link putting AI ethics into practice. arXiv preprint arXiv:2011.12750, pp. 1–22. https://arxiv.org/abs/2011.12750 (2020)
28. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**, 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
29. Haenlein, M., Kaplan, A.: A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. Calif. Manag. Rev. **61**(4), 5–14 (2019). https://doi.org/10.1177/0008125619864925
30. Hao, K.: We read the paper that forced Timnit Gebru out of Google. Here's what it says. MIT Technology Review. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/ (2020). Accessed 10 Feb 2022
31. Heaven, W.D.: The new version of GPT-3 is much better behaved (and should be less toxic). MIT Technology Review. https://www.technologyreview.com/2022/01/27/1044398/new-gpt3-openai-chatbot-language-model-ai-toxic-misinformation/ (2022). Accessed 9 Feb 2022
32. Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., Steinhardt. J.: Aligning AI with Shared Human Values. arXiv preprint arXiv:2008.02275. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–29. https://arxiv.org/abs/2008.02275 (2021)
33. Johnson, D.G., Verdicchio, M.: Reframing AI discourse. Mind. Mach. **27**, 575–590 (2017). https://doi.org/10.1007/s11023-017-9417-6
34. Kahn, J.: The sun is setting on A.I.'s Wild West. Fortune. https://fortune.com/2021/04/27/the-sun-is-setting-on-a-i-s-wild-west/ (2021)
35. Kreps, S., McCain, R.M., Brundage, M.: All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. J. Exp. Polit. Sci. **9**(1), 104–117 (2020). https://doi.org/10.1017/XPS.2020.37
36. LaGrandeur, K.: How safe is our reliance on AI, and should we regulate it? AI Ethics (2021). https://doi.org/10.1007/s43681-020-00010-7
37. Lucy, L., Bamman, D.: Gender and representation bias in GPT-3 generated stories. In: Proceedings of the 3rd Workshop on Narrative Understandings, pp. 48–55. https://www.aclweb.org/anthology/2021.nuse-1.5/ (2021). Accessed 20 May 2021
38. McCormack, J., Gifford, T., Hutchings, P.: Autonomy, authenticity, authorship and intention in computer generated art. In: Ekárt, A., Liapis, A., Pena, M.L.C. (eds.) Computational Intelligence in Music, Sound, Art and Design, pp. 35–50. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-16667-0_3
39. McGuffie, K., Newhouse, A.: The radicalization risks of GPT-3 and advanced neural language models. arXiv preprint arXiv:2009.06807, pp. 1–12. https://arxiv.org/abs/2009.06807 (2020)
40. McNamara, A., Smith, J., Murphy-Hill, E.: Does ACM's code of ethics change ethical decision making in software development?" In: Proceedings of the 2018 26th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2018), November 4–9, 2018, Lake Buena Vista, FL, USA. ACM, New York, NY, USA: 1–7. https://doi.org/10.1145/3236024.3264833 (2018)
41. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. arXiv preprint arXiv:1908.09635, pp. 1–31. https://arxiv.org/abs/1908.09635 (2019)
42. Mezentsev, S.: Technological determinism: breakthrough into the future. european proceedings of social and behavioural sciences. In: Proceedings of the 11th International and Theoretical

Conference “Communicative Strategies of Information Society”, pp. 240–248. https://doi.org/10.15405/epsbs.2020.03.02.29. (2019)

43. Mikalef, P., Conboy, K., Lundström, J.E., Popovič, A.: Thinking responsibly about responsible AI and ‘the dark side’ of AI. Eur. J. Inf. Syst. Edit. (2022). https://doi.org/10.1080/0960085X.2022.2026621
44. Mikalef, P., Gupta, M.: Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. Inf. Manag. **58**(103434), 1–20 (2021). https://doi.org/10.1016/j.im.2021.103434
45. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**, 2141–2168 (2020). https://doi.org/10.1126/science.132.3429.741
46. Müller, V.C.: Ethics of artificial intelligence and robotics. In: Zalta E.N. (ed.) The Stanford Encyclopedia of Philosophy (Winter 2020 Edition). https://plato.stanford.edu/archives/win2020/entries/ethics-ai/ (2020). Accessed 20 May 2021
47. Nallur, V., Lloyd, M., Pearson, S.: Automation: an essential component of ethical AI?. arXiv preprint arXiv:2103.15739. In: Proceedings of the 15th Multi Conference on Computer Science and Information Systems, 20–23 July 2021, pp. 1–4. https://arxiv.org/abs/2103.15739 (2021)
48. Noble, S.U.: Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press, New York (2018). https://doi.org/10.18574/9781479833641
49. OpenAI. *Aligning Language Models to Follow Instructions.* https://openai.com/blog/instruction-following/ (2022). Accessed 10 Feb 2022
50. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, G., Mishkin, p. , Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. pp 1–68, arXiv preprint arXiv: 2202.02155, https://arxiv.org/abs/2203.02155. (2022)
51. Russel, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 4th edn. Pearson, Hoboken (2021)
52. Susser, D., Roessler, B., Nissenbaum, H.: Technology, autonomy, and manipulation. Internet Policy Rev. **8**(2), 1–22 (2019). https://doi.org/10.14763/2019.2.1410
53. Thierer, A.: The pacing problem, the Collingridge dilemma, and technological determinism. The Technology Liberation. https://techliberation.com/2018/08/16/the-pacing-problem-the-collingridge-dilemma-technological-determinism/ (2018). Accessed 10 Feb 2022
54. Van de Poel, I.: Embedding Values in Artificial Intelligence (AI) Systems. Mind. Mach. **30**(3), 385–409 (2020)
55. Van den Hoven, J., Vermaas, P.E., van de Poel, I. (eds.): Handbook of Ethics, Values, and Technological design: Sources, Theory, Values and Application Domains. Springer, Cham (2015)
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, https://arxiv.org/abs/1706.03762 (2017)
57. Walker, B., Soule, S.A.: Changing Company Culture Requires a Movement, Not a Mandate. Harvard Business Review. https://hbr.org/2017/06/changing-company-culture-requires-a-movement-not-a-mandate (2017). Accessed 13 Feb 2022
58. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., Schwartz, J.: AI Now Report 2018, pp. 1–62 (2018)
59. Zhu, L., Xu, X., Lu, Q., Governatori, G., Whittle, J.: AI and ethics—operationalizing responsible. In: Chen, F., Zhou, J. (eds.) Humanity Driven AI. Springer, Cham (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.