LLMs beyond the lab
the ethics and epistemics of real-world AI research

Mollen, Joost

DOI
10.1007/s10676-024-09819-w

Publication date
2025

Document Version
Final published version

Published in
Ethics and Information Technology

Ethics and Information Technology (2025) 27:6
https://doi.org/10.1007/s10676-024-09819-w

ORIGINAL PAPER

LLMs beyond the lab: the ethics and epistemics of real-world AI research

Joost Mollen[1]

Abstract
Research under real-world conditions is crucial to the development and deployment of robust AI systems. Exposing large language models to complex use settings yields knowledge about their performance and impact, which cannot be obtained under controlled laboratory conditions or through anticipatory methods. This epistemic need for real-world research is exacerbated by large-language models' opaque internal operations and potential for emergent behavior. However, despite its epistemic value and widespread application, the ethics of real-world AI research has received little scholarly attention. To address this gap, this paper provides an analysis of real-world research with LLMs and generative AI, assessing both its epistemic value and ethical concerns such as the potential for interpersonal and societal research harms, the increased privatization of AI learning, and the unjust distribution of benefits and risks. This paper discusses these concerns alongside four moral principles influencing research ethics standards: non-maleficence, beneficence, respect for autonomy, and distributive justice. I argue that real-world AI research faces challenges in meeting these principles and that these challenges are exacerbated by absent or imperfect current ethical governance. Finally, I chart two distinct but compatible ways forward: through ethical compliance and regulation and through moral education and cultivation.

Keywords Large language models · LLMs · Generative AI · Research ethics · Real-world research · Emerging technology · AI ethics · AI governance

Introduction

In March 2023, the Future of Life Institute released an open letter titled 'Pause Giant AI Experiments,' signed by a long list of prominent figures in artificial intelligence research and governance (Future of Life 2023). Prompted by recent developments in the capacities and public deployment of generative AI systems, the letter posited that AI labs were locked in an uncoordinated race to develop and release powerful AI systems into society even though the societal consequences of these technologies were unknowable, uncontrollable, and potentially disastrous. As a solution, the letter urged AI labs to immediately pause the development and training of large language models (LLMs) more powerful than the GPT-4 model for at least six months to understand the systems better, focus on implementing safety protocols for AI design, and develop robust AI governance systems to ensure the safety of powerful AI systems (Future of Life 2023; 2023b). While a pause never materialized, the research and development of large language models has not slowed down since.

This paper focuses on a type of AI research and development that has yet to receive much philosophical attention: research conducted under real-world conditions[1]. Aside

✉ Joost Mollen
j.k.mollen@tudelft.nl

[1] Department of Values, Technology, and Innovation, Delft University of Technology, Delft, The Netherlands

[1] For the purposes of this paper, I am not concerned with the idea of technology as a social experiment. In recent decades, sociologists and philosophers of science have increasingly conceptualized technology (or its introduction in society) as an experiment of sorts, for example, by framing it society as a laboratory (Krohn & Weyer, 1994), a real-world experiment (Gross, 2018; David & Gross, 2019), social experiment (Martin & Schinzinger, 1983; Van de Poel, 2013; Poel, 2016, 2017a, b) or a collective experiment (Latour, 2004; Stilgoe, 2016). The intention of this frame is to draw attention to the inherent uncertainty involved in the introduction of experimental technologies in society and stress that we *should* learn from this uncertainty (Van de Poel, 2013; Stilgoe, 2016). These accounts have also drawn criticism, for example, that the frame of technology-as-social-experiment is essentially irrelevant (Peterson, 2013; 2017) or that the concept of

Published online: 19 December 2024

from controlled laboratory studies, AI systems are routinely tested in their potential use setting. These real-world environments are also referred to as 'the wild' or 'everyday social contexts' (David & Gross, 2019, p.992). This real-world research is central to developing and deploying robust AI systems. Exposing AI systems to complex and unpredictable socio-technical environments can yield insights about their performance, which cannot be obtained under controlled laboratory conditions. Real-world research can differ from scientific research in that it, for example, does not necessarily employ experimental control techniques such as randomization, control groups, and isolating variables (Ansell & Bartenberger, 2016, 2017) or aim to accept or reject a particular hypothesis (Popper 2013; Rheinberger, 1997). Instead, real-world research is broadly more concerned with innovation, group or cluster-level interventions, realizing a desirable state of affairs or contextual success ('making the technology work in its context') and is often characterized by their absence of control to retain the 'natural' representative quality of the research environment (Ansell & Bartenberger, 2016, 2017).

This lack of attention is problematic for at least two reasons. First, as mentioned, real-world AI research is widespread and crucial to AI development and deployment. With the intent to promote innovation, real-world AI research is routinely enabled and encouraged by 'soft law' mechanisms such as regulatory sandboxes (Ranchordas, 2021) or made exempt from many regulatory demands in AI governance regulations such as the European Union's AI Act (Colonna, 2023). Second, real-world research raises ethical concerns. While there has been increasing scholarly and political attention to the ethics and governance of generative AI systems – such as their capacity to violate copyright laws (Lucchi, 2023), create biased output (Zhou et al., 2024), enable plagiarism (Kwon 2024) or manipulation (Klenk, 2024) and cause ecological impact (Bender et al., 2021) – similar attention has not been extended to *researching* generative AI systems. However, scholars are increasingly drawing attention to the ethical issues that real-world research with emerging technologies brings about. These issues include the avoidance of democratic accountability by investigators (Taylor, 2021), causing physical harm (Stilgoe, 2020; Colonna, 2023), violating human rights (Amnesty, 2020), the imposition of 'dominating' risk (Maheshwari & Nyholm, 2022), and the unequal ethical demands between various categories of real-world research (Mollen, 2024).

In order to address this gap, this paper provides an analysis of real-world research with generative AI systems and the large language models on which they are built. I will assess both its epistemic value and ethical dimensions. First, I outline the epistemic need for real-world research with large language models. I discuss the limitations of controlled or anticipatory learning methods such as laboratory benchmarking and forecasting and argue that these limitations are exacerbated by large-language models' opaque internal operations and potential for emergent behavior. Second, I argue that this creates an epistemic need to acquire knowledge about large language models through real-world research and outline various potential learning outcomes. Third, I argue that despite its epistemic value, real-world research with AI brings about various ethical concerns that must be taken seriously. I structure these concerns alongside four moral principles that have influenced research ethics standards: non-maleficence, beneficence, respect for autonomy, and (distributive) justice. I then argue that these moral concerns are exacerbated by absent or imperfect current ethical governance. Finally, I discuss two distinct but compatible ways forward regarding embedding research ethics in real-world AI research: through ethical compliance and regulation and through moral education and cultivation.

The limits of controlled and anticipatory learning about large language models

In this section, I will discuss the limitations of learning methods that allow us to gather knowledge about large language models before they are studied under real-world conditions. I will discuss benchmarking and forecasting. In a nutshell, the shortcoming of these methods is that they either rely on what can be currently known about the model in a controlled and unrepresentative context or rely on anticipatory or predictive information, which is speculative to a certain degree. Specifically, I argue that these shortcomings are exacerbated by large-language systems' largely opaque internal operations and potential for emergent behavior.

The limits of benchmarking

Benchmark tests are standardized software performance tests that measure a system's performance across various tasks and topics. Benchmarking allows the evaluation of the quality of the systems or models and the ability to compare this to the performance of other AI systems. One example of a language model benchmark is Stanford's Holistic Evaluation of Language Models (HELM) (Liang et al., 2022; Bommasani et al., 2023). HELM involves a multi-metric evaluation of a language model across various scenarios and metrics. These scenarios can involve, for example, answering questions ranging from mathematics to ethics, as well as summarization and information retrieval. Metrics

experimentation is stretched out to a point where it loses any analytical value (Karvonen & van Heur, 2014; Huitema et al., 2018; Hansson, 2019, notes 1).

include, among others, fairness, accuracy, bias, robustness, and toxicity (Liang et al., 2022; Bommasani et al., 2023). Benchmark tests can allow for transparent communication to users, regulators, and the larger public about the quality of specific models across various scenarios and metrics and indicate the need to amend the model if low-performance scores are measured.

However, benchmark tests conducted under laboratory conditions face various limitations. First, benchmarks can run into the potential problem of restricted scope, in that tests might target only known capabilities and overlook unknown capabilities. Second, there is the problem of external validity –can the result be transported outside the research context? It can prove difficult to accurately model the conditions and interactions a large language model might be subject to when embedded in a more extensive socio-technical system (Srivastava et al., 2022). Third, there is a problem of potential construct validity: the degree to which a test captures what it aims to assess. For example, particular LLM benchmarks aim to capture normative concepts, such as fairness or safety, yet lack clear philosophical foundations. Fourth, large language models can bring about risks and social consequences – such as the automation of jobs – which cannot be measured at the technology level (Mökander & Floridi, 2021).

The limits of forecasting

A form of anticipatory learning about large language models is through various foresight approaches (Brey, 2017). These approaches aim, as Brey notes, to "project likely, plausible or possible future products, applications, uses and impacts that may result from the further development and introduction of an emerging technology" into society based on what are inherent or necessary system features or conditions for their realization (Brey, 2017). One example is the Delphi method – an anticipatory technique that establishes expert consensus on current and potential future developments on a particular issue. A recent study employed this method to study the possible impact of large language models on scientific practice (Fecher et al., 2023). Similar anticipatory studies have stressed the social impact of large language models on medical research and care (Clusmann et al., 2023), the labor market (Eloundou et al., 2023), mental health services (Van Heerden et al., 2023), and crime (Europol, 2023), among others.

However, forecasting methods are limited since the complex socio-technical environments that these models aim to operate within make predictions with a high degree of confidence difficult. There exists disagreement as to the degree to which this shortcoming of forecasting methods can eventually be resolved and, hence, whether the inability to accurately predict the trajectory of a technology is a mere methodological obstacle or an ontological limit (Liebert & Schmidt, 2010). This problem is central to the Collingridge dilemma that states that we have the most control to shape (the trajectory of a) technology when there is little knowledge about its social impact – and vice versa (Collingridge, 1982; Kudina & Verbeek, 2019). Additionally, Van de Poel argues that forecasting might focus disproportionately on tantalizing but unlikely scenarios and consequently draw attention away from more realistic but less thought-provoking issues that need attention more (Van de Poel, 2016). In the context of large-language models, we might group concerns about machine superintelligence in this corner.

Exacerbating limitations: system opaqueness and emergent behaviour

To some extent, the shortcomings mentioned above are the case for every technology. However, they do not necessarily apply to *the same degree* for every technology. I argue that large language models have additional characteristics that make controlled and anticipatory learning more difficult than other technologies: they are opaque technologies and (potentially) capable of emergent behavior. These two features – the opaque nature of large language models and their potential for emergent behavior – further trouble attempts to understand both a system's current and future capacity and behavior.

First, large-language models are *opaque* technologies. With opacity, I refer to the idea that we have limited access to explanations about an artificial system's inner workings or reasonings (Smith, 2021; Vaassen, 2022). Burrell distinguishes between three sources of opacity: either through an (intentional) failure of corporate or state communication, a lack of expertise or technical literacy, or due to the system's inherent features and required scale of use (Burrell, 2016). The latter source is relevant to my point. To take OpenAI's GPT large language model as an example, the number of parameters of GPT-1 grew from about 117 million parameters in 2018 (Hadi et al., 2023) to 1.5 billion (GPT-2) to 175 billion parameters for GPT-3 (Zhang and Li 2019). Additionally, large language models are trained on massive datasets, making it often difficult to understand the exact makeup of the training data (Bender et al., 2021). The opacity induced by this scale makes fully understanding the current and future behavior of a powerful large-language model difficult.

A second feature of large-language models that might contribute to limited anticipatory learning about a system is the possibility of emergent behavior (Wei et al., 2022; Hagendorff, 2023; Webb et al., 2023). This refers to the idea that, due to the scale of the models involved and their complex

internal interactions, a large language model can produce unpredictable behavior that the system was not necessarily trained for and was absent in smaller model iterations. In other words, the increase in scale produces qualitative new behavior. This presents a problem for extrapolating the capabilities of a larger language model based on the capacities of a smaller version since additional scaling could further expand the capabilities of a model (Wei et al., 2022). While empirical data of potential emergent behavior has been collected, as Srivastava and colleagues note, "we are unable to reliably predict the scale at which new breakthroughs will happen" and might "be unaware of additional breakthroughs that have already occurred but not yet been noticed experimentally" (2022). Additionally, Hagendorff claims that traditional benchmark tests cannot detect emergent abilities (2023). Whether this behavior is actually 'emergent,' in the sense that scale causes fundamental changes in the model's behavior, is a current matter of debate. Others have argued what some label as emergent behavior is better explained through other means, such as metric choices or in-context learning (Schaeffer et al., 2023; Hodel & West, 2023; Lu et al., 2023). Regardless of the origin of those capacities or what we decide to label as emergent behavior, for my purposes, the point stands that there are difficulties in gaining knowledge about the total range of capacities of a largelanguage model.

So, while controlled and anticipatory methods might teach us how powerful large language models operate under specific controlled conditions, they provide us with little operational understanding and confidence in how the generative AI system might perform under real-world conditions. Here, an epistemic need emerges. In the next section, I discuss the specific learning outcomes that real-world research can offer.

The epistemic value of real-world AI research

In this section, I discuss the epistemic value of real-world AI research. Since controlled and anticipatory learning is limited, this creates an epistemic need to acquire knowledge about AI systems through research under real-world conditions. Exposing AI systems to diverse, representative, and unpredictable environments can yield insights about their performance, which are impossible or difficult to obtain under anticipatory laboratory conditions.

First, real-world AI research can show how a particular large language model performs in its potential use setting rather than in a controlled research setting. For example, New York City Public School's AI Policy Lab tested how large language models can aid educational tasks such as lesson planning (GovTech 2023). The British Department of Education researched whether ChatGPT could aid officials in summarizing and comparing various training plans (Seddon 2023). Other examples include studies that have explored the impact of LLMs on the development of critical thinking skills in high school classes level (Bitzenbauer, 2023) and their potential to identify errors in student homework and provide them with personalized feedback to streamline the assessment procedure (Bewersdorff et al., 2023).

Second, real-world research allows the possibility to discover whether a large language model is *comparatively* superior or inferior to another in a specific use context and, thus, which model better suits a particular socio-technical environment. For example, the U.S. Department of Defence conducted tests with five different large language models to study to what degree they could improve access times to internal information or even help plan responses to potential global conflicts (Manson 2023). Another recent study compared the performance of various large language models - ChatGPT, Bard, Claude, and ChatLlama – on identifying phishing emails (Heiding et al., 2023). Alternatively, the performance of a large language model can be compared in a given context to that of human actors. The same study tested whether phishing emails made by humans, large language models, or a combination of the two were more successful in convincing a subject pool of 112 Harvard students, with human-written emails far outperforming the AI-generated content (Heiding et al., 2023).

Third, real-world research allows learning about how generative AI can be successfully embedded within specific institutions. The successful embedding of a novel technology within an organization often goes beyond mere technical capacity but largely depends on social factors. Thus, real-world research allows for learning about, for example, which protocols or normative frameworks can best guide a responsible and effective use of the AI system or what additions might be necessary to secure responsible embedding, such as digital watermarks to algorithmically identify AI-generated content (Kirchenbauer et al., 2023). Real-world research could thus offer social learning about the successful and responsible embedding of generative AI systems within operations.

Fourth, real-world research allows for monitoring and responding to emergent social impacts of large language models. For example, the EU's Artificial Intelligence Act (AIA) mandates that the providers of high-risk AI systems must engage with post-marketing surveillance to monitor, document, and analyze the performance of these systems throughout their life cycle (Mökander et al., 2022). Post-market surveillance refers to a set of obligatory monitoring activities a manufacturer has to perform to ensure the performance and safety of their product *after* it has been released on the market (Pane et al., 2019; Beckers et al.,

2021). During post-marketing surveillance, providers are expected to report serious malfunctions and take immediate action to either correct this malfunction to bring the system back in conformity or withdraw it from the market (Mökander et al., 2022). Through these measures, the performance and continued safety of these products can be closely monitored and, ideally, withdrawn from the market in the case of negative social consequences.

Fifth, real-world research provides an opportunity to learn about a generative AI system's normative and moral consequences (Van de Poel, 2017b). Real-world research offers the chance to see whether a system meets ethical requirements, for example, as part of an ethics-by-design approach (Brey & Dainow, 2023). The Dutch government's Impact Assessment Fundamental Rights and Algorithms notes that real-world test beds can help identify harms to fundamental rights before such models are publicly released (Janssen, 2020; Ministerie van Algemene Zaken 2022). Harbers and Overdiek have also argued that real-world living labs could contribute to ethical AI design, development, and deployment (Harbers & Overdiek, 2022). Mökander and colleagues have recently proposed 'ethics-based auditing,' which assesses large language models to determine their consistency with relevant moral values (Mökander & Floridi, 2021).

Finally, Van de Poel has argued that since research environments are 'small-scale' compared to a monitored public-wide market release, potential negative consequences will be comparatively more minor, less costly, and more likely to be possible to amend (Van de Poel, 2017b). Costly refers to the scale or amount of negative consequences and the resources necessary to resolve or amend them. Van de Poel argues that while real-world research would potentially still be more 'costly' than anticipatory strategies (as in that negative social consequences might occur that would not have happened if they were not researched under real-world conditions in the first place) since technologies in a real-world test environment or under post-marketing surveillance are closely monitored, we will know at an early stage when negative consequences occur and can quickly feed this information back into improving either the design or embedding process (Van de Poel, 2017a; Poel, 2017b).

Thus, real-world AI research meets a critical epistemic need since it can provide valuable insights into the successful development and embedding of generative AI systems that we cannot acquire through controlled or anticipatory methods. However, this learning also raises ethical concerns, which I will outline in the next section.

The ethics of real-world AI research

In this section, I discuss various ethical dimensions of real-world research with generative AI and large language models. Despite its epistemic value, real-world AI research also raises ethical concerns. I organize these ethical concerns along four moral principles that underpin many legal, professional, and moral standards regarding ethical research: non-maleficence, beneficence, justice, and respect for personal autonomy (Beauchamp & Childress, 1994). These moral principles have been influential in shaping much of contemporary research ethics and, additionally, the moral evaluation of introducing experimental technology into society (Van de Poel, 2016) and AI ethics guidelines such as the EU's Ethics Guidelines for Trustworthy Artificial Intelligence or OECD's Recommendation of the Council on Artificial Intelligence (Nikolinakos, 2023; Porter et al., 2024). I aim not to defend or criticize this framework or particular interpretations of the moral principles involved or argue that this list intends to be complete. Instead, I use these moral principles to capture and organize a wide range of relevant ethical issues in real-world AI research and discuss how real-world AI research might bring about context-specific challenges in addressing these issues.

Non-maleficence

First, the moral principle of non-maleficence refers to the idea that research interventions should 'do no harm' (Van de Poel, 2016). Here, I define harm not merely in a physical sense but as any wrongful setback to, or thwarting of, an interest, such as the violation of a right (Feinberg, 1984). Researchers are obligated to not cause harm or prevent harm from arising as a consequence of the research intervention.

The risks that real-world AI research might bring about can vary. For example, Colonna has argued that testing artificial intelligence under real-world conditions can present "risks to individual's health, safety and fundamental rights, as well as broader societal concerns" (Colonna, 2023, p.28). An example of such a broader societal concern is the environmental impact of AI systems. Due to the energy consumption and global resources required during the entire lifespan of an AI system, scholars have increasingly drawn attention to the carbon cost and environmental impact of AI systems (Dhar, 2020; Bender et al., 2021). Hence, the (real-world) research and development of powerful large language models - given their current energy consumption - will further impact the environment, increase the carbon footprint, and contribute negatively towards mitigating climate change (Dobbe & Whittaker, 2019; McDonald et al., 2022; Lakim et al., 2022; Rillig et al., 2023). This means that even if the potential negative consequences in a real-world research

setting will be comparatively minor, generative AI systems or large language models can still carry risks, some of which may be substantial. When researching these systems on a group level, we effectively expose populations interacting with these systems to these risks.

Real-world AI research, however, poses challenges to prevent or mitigate harm for at least two reasons. First, when research is conducted within a real-world environment, predicting, containing, and identifying risks - or even identifying which persons might be affected by the intervention can become more difficult due to the interconnected and complex real-world environments in which some AI systems are tested. If researchers cannot identify who is harmed during or after the experiment, compensating or remedying harm becomes difficult or impossible.

Second, it is unclear how early detection of negative consequences might lead to adjustments to the design or implementation of large language models. As mentioned above, Van de Poel has argued that one of the benefits of learning about technology through closely monitored small-scale introduction is that, ideally speaking, we will know at an early stage when negative consequences occur and can quickly feed this information back into improving either the design or embedding process (Van de Poel, 2017a; Poel, 2017b). This idea of controlled, iterative learning also underpins much of post-marketing monitoring and regulatory sandboxes. However, large language models are complex digital technologies. Unlike physical devices, such as toasters or cars, that can be redesigned in response to specific safety concerns, large language models are complex, adaptive systems that do not allow for straightforward design modifications in response to individual adverse outcomes.[2] At best, monitoring might prompt a recall of a particular technology. In some cases – see some of the examples in Sect. 3 – parties testing out a particular large language model only have (paid) access to use the model and are not able to make changes to the underlying model when negative consequences might arise. Instead, they only have the power to decide how they will use the model or whether they will use it at all. Hence, even if negative consequences arise in a real-world test, this does not necessarily mean that these insights will be translated back into fundamental changes to the models.

Beneficence

Second, the moral principle of beneficence prescribes that aside from avoiding harm, researchers are also obligated to 'do good,' for example, by producing social value, doing more good than harm, or removing existing harms in the world (Van de Poel, 2016). If real-world AI research brings about risks or harms to particular persons or groups, such research should at the very least be conducted with the intention – and under the reasonable belief – that it will bring social value into the world, either by directly benefitting people's lives or, for example, by lowering the demands on public resources through more efficient operations.

One potential challenge to this aim of beneficent real-world AI research is the increased privatization of AI research. In recent years, the center of gravity of AI research and development has increasingly shifted away from (public) academic institutions to private companies (Jurowetzki et al., 2021; Giziński et al., 2024). As the *who* of AI research transitions towards industry, this changes *what* is being learned and *who* has the power to decide what is being learned. The AI industry plays a large role in identifying, influencing, and shaping the 'problems' that receive research focus and funding (Khanal et al., 2024). Consequently, private interests can constrain research scope or funding and limit research topics not in line with the corporate interest but which might be socially relevant (Jurowetzki et al., 2021). This way, corporate interests set the AI research agenda, which might not necessarily align with societal goals. For example, industry-driven learning might favor short-term monetization and competitive advantages and hold lower expectations for the social value of their research or other considerations such as environmental costs, societal externalities, and ethical challenges (Bender et al., 2021; Jurowetzki et al., 2021). This, Jurowetzki and colleagues argue, "bolsters the case for increasing AI research capabilities in academia and government in order to ensure that public interests can continue playing an active role in monitoring and shaping the trajectory of powerful AI systems" (2021, p.2).

Respect for autonomy

Respect for autonomy refers to the obligations of researchers to protect and secure the autonomy of persons or groups involved in the research (Van de Poel, 2016). Persons have a right to make autonomous decisions in that they should have control over their own lives, bodies, and data and make decisions about them according to their reasons, motives, and interests. Since research can intervene within a person's sphere of autonomy, particular research ethics mechanisms, such as informed consent and withdrawal procedures, aim to help safeguard a person's autonomy.

Here, real-world AI research raises various ethical concerns. First, there is a question of availability and access to information about the research. Since real-world AI research takes place in 'natural' environments, people might not be aware that they are part of a research project without being adequately informed. If people are unaware that they are

[2] I want to thank an anonymous reviewer for stressing this point.

part of a research project, they cannot make an informed decision to participate in the research and thereby consent to its potential associated risks and benefits. However, even if a person is aware of the research happening, issues arise regarding the ability to opt-out. For example, how can a person meaningfully opt-out from interacting with a generative AI system that is tested in an area that is difficult or costly to avoid, such as a place of work or government institutions? Additionally, there are questions regarding data ownership. How can subjects exposed to real-world AI research keep control over their data (mainly when industry parties might conduct such research), and what rights and abilities do they have to amend or withdraw their data after the fact?

Distributive justice

The principle of justice in research ethics generally refers to researchers' obligations relating to distributive justice, i.e., the just distribution of the research's benefits and risks. This includes concerns regarding the equitable selection of subjects, avoiding exploitation, protecting vulnerable subjects, or ensuring that research in which vulnerable subjects participate is beneficial to them (Van de Poel, 2016).

Real-world AI research can bring about various issues of distributive justice. Due to a lack of ethical governance (I will expand on this in the next section), there may be a tendency to conduct research in areas or regions with less regulatory oversight or among individuals or groups who lack sufficient awareness of these risks. This would mean that risks are disproportionately placed on those communities that enjoy the least protection. Additionally, it might be difficult to provide safeguards for vulnerable persons or groups when these persons or groups are challenging to identify in a real-world setting. If researchers are not aware of the exact demographic makeup of their subject pool, it will be difficult to exclude – or award additional protections to - vulnerable individual subjects or groups.

Another question concerns how affected people and groups can share in the benefits of real-world AI research that is subjecting them to particular risks. Here, the issue of increased AI privatization also plays a role in *who* benefits from this learning. As mentioned, knowledge about AI systems or their performance is increasingly concentrated within private companies. This data could be difficult or undesirable to share with academia for proprietary reasons or market advantage (Jurowetzki et al., 2021) and thus difficult to reproduce and replicate (Giziński et al., 2024) or made subject to independent ethical scrutiny (Resseguier & Ufert, 2024). Even when public institutions run their own tests with embedding particular instances of generative AI, such tests can still benefit corporate interest if those public institutions use AI systems developed by industry and firmly embed them within their operations, potentially leading to a lock-in problem.

Lacking ethical governance of real-world AI research

So far, I have described a tension between the epistemic value of real-world AI research and various ethical concerns this type of research can bring about. This prompts questions regarding the need for external scrutiny. In this section, I discuss that navigating these tensions is difficult due to an absent or imperfect scope of ethical governance, which exacerbates the abovementioned problems.

Generally, research ethics governance mechanisms – such as guidelines, protocols, or ethical review boards or committees – aim to address or (help) navigate the ethical tensions described in the previous section. They can do so by providing action-guiding norms or through various research ethics mechanisms that provide a means of reviewing research proposals (and their research's risks and risk mitigation strategies) and holding researchers accountable for research malpractice and subject redress.

However, research under real-world conditions with generative AI is conducted in a space lacking ethical standards and protocols (Mollen, 2024). While clear research ethical demands generally bind scientific research, such mechanisms are often absent in research conducted by industry or public parties. While there has been an increase in AI guidelines and ethics codes, Munn has argued that these ethical principles are largely useless and do not impact practice since they are 'meaningless' (contested or incoherent), 'isolated' (applied to domains that ignore ethics), and 'toothless' (without consequence or in-line with industry interest) (Munn, 2023, p. 872). This leaves people and groups vulnerable since there are no mechanisms for external scrutiny, and people are not effectively given control to counter-act experimental impositions.

The absence of research ethics governance can also enable the evasion of ethical demands elsewhere. When different ethical demands are placed on two research domains, one research domain can avoid such demands by placing particular research activities outside the scope of the demands they are subject to (Metcalf & Crawford, 2016; Colonna, 2023; Mollen, 2024). For example, while specific data might not be captured without the subjects' consent by scientific research, when no such demands are placed on corporate researchers, the latter could collect this data. At that point, it becomes public data that can be used. In this way, the absence of research ethics governance in one domain can come at the expense of those whom other ethical demands aim to protect.

Even if such research is conducted by parties operating within an ethically regulated domain – for example, scientific publicly-funded research – the available ethical guidelines or protocols might not help address researchers' moral and regulatory challenges. For example, AI and data scholars have increasingly called for research ethics reforms to address current limitations (Vitak et al., 2017; Raymond, 2019). Resseguier and Ufert, for example, have argued in favor of three adaptions of current research ethics standards and mechanisms to better asses scientific AI research (2023). First, existing research ethics frameworks need to adapt and move beyond a sole focus on protecting individual human participants when identifying and mitigating AI risk and include an assessment of risks and harms to communities, society at large, and the environment. Second, the period when risks and harms are considered needs to be extended from the research stage to when the AI system is deployed (2023). Third, Resseguier and Ufert argue that much of the data that fuels current AI research comes from scraping existing data, using existing data sets, or collecting data through in-direct methods such as public sensors. Under current research ethical guidelines, this data is often considered exempt from ethical review (Ada Lovelace Institute, 2022, p.5). While this data might be innocuous in the original study, it can be re-combined to create invasive datasets (Metcalf & Crawford, 2016). Hence, adapted research ethics for AI research needs to be sensitive to this kind of data collection.

However, as long as research under real-world conditions is conducted in partnership with parties not bound by these ethical demands – such as many industry parties – these research ethics reforms only target scientific AI research at best. Real-world research with artificial intelligence often involves research collaborations between private, public, and knowledge institutions (Ada Lovelace Institute, 2022, p.7). Such collaborations between different stakeholders can thus cause confusion about how (moral) responsibilities should be divided and how particular ethical concerns can be navigated or resolved in the case of conflicting values or interests within the research consortium.

Moving forward: embedding ethics within real-world AI research

The above section presents a persuasive case to ameliorate the current situation in which much of real-world AI research is conducted under imperfect ethical governance – or in its complete absence. In this section, I will briefly discuss the benefits and drawbacks of two distinct but mutually compatible approaches to embedding research ethics within real-world AI research: through ethical compliance and regulation and through moral education and cultivation.

The first approach relies on ethical governance through regulation, such as mandatory ethical compliance or an institutional review board review. One example of this approach is the conditions the EU's AI Act places on research with high-risk AI outside the scope of regulatory sandboxes. These include requiring informed consent, additional protections for vulnerable populations, the protection of personal data, removing personal data after persons have withdrawn their consent, outlining the roles and responsibilities of all parties involved, and creating a real-world testing plan detailing the goals and duration of the research which needs to be registered in a EU-wide database and submitted to 'competent market surveillance authorities' (AI Act 72b).

A benefit of such ethical governance is that it is mandatory, creating a concrete incentive to industry and public parties aiming to research a particular AI system under real-world conditions. Additionally, it provides governments with a 'check-point' to assess and influence what kind of research is conducted with (generative) AI under real-world conditions, ensuring that the research creates public value. On the other hand, mandatory regulations can also bring about a 'checklist' ethics mentality, creating additional costs and demands for government oversight agencies and practical and conceptual challenges to meeting these demands when conducting research under real-world conditions, for example, difficulties in obtaining informed consent or protecting vulnerable groups when it is difficult to identify research subjects.

An alternative approach aims to foster ethical AI research through non-mandatory incentive structures such as independent review boards providing research ethics advice, ethical guidelines, workshops, design activities, games, or roleplaying for practitioners to create increased awareness about the moral dimensions of their research practices (Wong et al. 2020), ethics training and cultivating virtues (Hagendorff, 2022), structured discussions throughout the research process for ethical reflection, conference and journal standards (Polonioli et al., 2023), etc. One example would be the Dutch Fundamental Rights and Algorithm Impact Assessment (FRAIA), a human rights dialogue and reflection tool for developers or deployers of algorithmic systems.

A benefit of this approach is that it aims to motivate, interest, and cultivate a researcher's conviction to do good rather than to be merely compliant with mandatory regulation. However, an apparent shortcoming of this approach is its largely self-regulating nature, meaning that if these approaches are unsuccessful or purposefully neglected, they leave little protection for those affected by the research intervention.

Conclusion

In this paper, I have discussed the epistemic value of real-world AI research and the ethical concerns this type of research brings about. While generative AI and large language models hold great promise, it is important that they are developed in a manner that is ethical and consistent with moral principles. While there is a clear epistemic need for real-world AI research – exacerbated by large-language models' opaque internal operations and potential for emergent behavior – this does not mean this research should be conducted without the ethical guardrails we find in other types of (scientific) research. Currently, real-world AI research is conducted in a space that lacks proper ethical governance, leaving persons and groups without due protection and exacerbating real-world AI research's moral concerns. Hence, we should strive to ameliorate the current situation by drawing from two distinct but mutually compatible approaches to embedding research ethics within real-world AI research: ethical compliance and regulation and moral education and cultivation. While these methods might have their respective downsides, a balanced approach to incorporating ethics in real-world AI research is not only necessary but overdue.

Funding No funds, grants, or other support was received. The PhD research project of the author is made possible by The Province of South Holland, The Netherlands. The author has no relevant financial or non-financial interests to disclose.

References

Ada Lovelace Institute (2022). *Looking before we leap: Ethical review processes for AI and data science research*. https://www.adalovelaceinstitute.org/report/looking-before-we-leap/

Amnesty International (2020). *We sense trouble: Automated discrimination and mass surveillance in predictive policing in the Netherlands*. Retrieved Nov 2024, from https://www.amnesty.org/en/documents/eur35/2971/2020/en

Ansell, C. K., & Bartenberger, M. (2016). Varieties of experimentalism. *Ecological Economics*, *130*, 64–73.

Ansell, C., & Bartenberger, M. (2017). The diversity of experimentation in the experimenting society. *New perspectives on Technology in Society* (pp. 36–58). Routledge.

Beauchamp, T. L., & Childress, J. F. (1994). *Principles of biomedical ethics*. Edicoes Loyola.

Beckers, R., Kwade, Z., & Zanca, F. (2021). The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. *Physica Medica*, *83*, 1–8. https://doi.org/10.1016/j.ejmp.2021.02.011

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). https://doi.org/10.1145/3442188.3445922

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, 100177. https://doi.org/10.1016/j.caeai.2023.100177

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, *15*(3), ep430. https://doi.org/10.30935/cedtech/13176

Bommasani, R., Liang, P., & Lee, T. (2023). Holistic evaluation of Language models. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.15007

Brey, P. (2017). Ethics of emerging technology. *The ethics of technology: Methods and approaches*, 175-191.

Brey, P., & Dainow, B. (2023). Ethics by design for artificial intelligence. *AI and Ethics*, 1–13.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J. N., Laleh, N. G., & Kather, J. N. (2023). The future landscape of large language models in medicine. *Communications Medicine*, *3*(1), 141. https://doi.org/10.1038/s43856-023-00370-1

Collingridge, D. (1982). The social control of technology.

Colonna, L. (2023). The AI act's research exemption: A mechanism for Regulatory Arbitrage? *YSEC Yearbook of Socio-Economic constitutions 2023: Law and the governance of Artificial Intelligence* (pp. 51–93). Springer Nature Switzerland.

David, M., & Gross, M. (2019). Futurizing politics and the sustainability of real-world experiments: What role for innovation and exnovation in the German energy transition? *Sustainability Science*, *14*, 991–1000.

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nat Mach Intell*, *2*(8), 423–425. https://doi.org/10.1038/s42256-020-0219-9

Dobbe, R., & Whittaker, M. (2019). AI and climate change: How they're connected, and what we can do about it. *AI Now Institute Medium*, *17*.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv Preprint arXiv:2303 10130*. https://doi.org/10.48550/arXiv.2303.10130

Europol (2023). *ChatGPT - The impact of Large Language Models on Law Enforcement*, a Tech Watch Flash Report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg.

Fecher, B., Hebing, M., Laufer, M., Pohle, J., & Sofsky, F. (2023). Friend or foe? Exploring the implications of large Language models on the Science System. *arXiv Preprint arXiv:2306 09928*. https://doi.org/10.48550/arXiv.2306.09928

Feinberg, J. (1984). *Harm to others* (Vol. 1). Oxford University Press.

Felt, U., Wynne, B., Callon, M., Gonçalves, M. E., Jasanoff, S., et al. (2007). *Taking European Knowledge Society seriously*. Directorate-General for Research, Science.

Future of Life Institute (2023). *Pause giant AI experiments: An open letter*. Retrieved from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Future of Life Institute (2023b). *Policymaking in the pause*. Future of Life Institute. Retrieved from https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf

Giziński, S., Kaczyńska, P., Ruczyński, H., Wiśnios, E., Pieliński, B., Biecek, P., & Sienkiewicz, J. (2024). Big tech influence over AI research revisited: Memetic analysis of attribution of ideas to affiliation. *Journal of Informetrics*, *18*(4), 101572.

GovTech (2023, October 17). *NYC schools working with experts to launch AI Policy Lab*. GovTech. https://www.govtech.com/education/k-12/nyc-schools-working-with-experts-to-launch-ai-policy-lab

Gross, M. (2018). Real-world experiments as generators of sociotechnical change. *Energy as a Sociotechnical Problem* (pp. 125–138). Routledge.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.

Hagendorff, T. (2022). A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology*, *35*(3), 55.

Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv Preprint arXiv:2303 13988*. https://doi.org/10.48550/arXiv.2303.13988

Hansson, S. O. (2019). Farmers' experiments and scientific methodology. *European Journal for Philosophy of Science*, *9*(3), 32.

Harbers, M., & Overdiek, A. (2022). Towards a living lab for responsible applied ai. In: Proceedings of the DRS 2022. Retrieved from https://doi.org/10.21606/drs.2022.422

Heiding, F., Schneier, B., Vishwanath, A., & Bernstein, J. (2023). Devising and detecting phishing: Large language models vs. smaller human models. *arXiv Preprint arXiv:2308 12287*. https://doi.org/10.48550/arXiv.2308.12287

Hodel, D., & West, J. (2023). Response: Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2308.16118*. https://doi.org/10.48550/arXiv.2308.16118

Huitema, D., Jordan, A., Munaretto, S., & Hildén, M. (2018). Policy experimentation: Core concepts, political dynamics, governance and impacts. *Policy Sciences*, *51*, 143–159.

Janssen, H. L. (2020). An approach for a fundamental rights impact assessment to automated decision-making. *International Data Privacy Law*, *10*(1), 76–106. https://doi.org/10.1093/idpl/ipz028

Jurowetzki, R., Hain, D., Mateos-Garcia, J., & Stathoulopoulos, K. (2021). The Privatization of AI Research (-ers): Causes and Potential Consequences--From university-industry interaction to public research brain-drain? *arXiv preprint arXiv:2102.01648*.

Karvonen, A., & Van Heur, B. (2014). Urban laboratories: Experiments in reworking cities. *International Journal of Urban and Regional Research*, *38*(2), 379–392.

Khanal, S., Zhang, H., & Taeihagh, A. (2024). Why and how is the power of Big Tech increasing in the policy process? The case of generative AI. *Policy and Society*, puae012.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. *arXiv Preprint arXiv:2301 10226*. https://doi.org/10.48550/arXiv.2301.10226

Klenk, M. (2024). Ethics of generative AI and manipulation: A design-oriented research agenda. *Ethics and Information Technology*, *26*(1), 9.

Krohn, W., & Weyer, J. (1994). Society as a laboratory: The social risks of experimental research. *Science and Public Policy*, *21*(3), 173–183. https://doi.org/10.1093/spp/21.3.173

Kudina, O., & Verbeek, P. P. (2019). Ethics from within: Google Glass, the Collingridge dilemma, and the mediated value of privacy. *Science Technology & Human Values*, *44*(2), 291–314. https://doi.org/10.1177/0162243918793711

Kwon, D. (2024, July 30). *Ai is complicating plagiarism. how should scientists respond?* Nature News. https://www.nature.com/articles/d41586-024-02371-z

Lakim, I., Almazrouei, E., Abualhaol, I., Debbah, M., & Launay, J. (2022, May). A holistic assessment of the carbon footprint of noor, a very large Arabic language model. In *Proceedings of BigScience Episode# 5--Workshop on Challenges & Perspectives in Creating Large Language Models* (pp. 84-94). https://doi.org/10.18653/v1/2022.bigscience-1.8

Latour, B. (2004). Which protocol for the new collective experiments. *Experimental Cultures*, 17–36.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., & Koreeda, Y. (2022). Holistic evaluation of language models. *arXiv Preprint arXiv:2211 09110*. https://doi.org/10.48550/arXiv.2211.09110

Liebert, W., & Schmidt, J. C. (2010). Towards a prospective technology assessment: Challenges and requirements for technology assessment in the age of technoscience. *Poiesis & Praxis*, *7*, 99–116.

Lu, S., Bigoulaeva, I., Sachdeva, R., MadabushiH T, & Gurevych, I. (2023). Are Emergent abilities in large Language models just In-Context learning? *arXiv Preprint arXiv:2309 01809*. https://doi.org/10.48550/arXiv.2309.01809

Lucchi, N. (2023). ChatGPT: A case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, 1–23.

Maheshwari, K., & Nyholm, S. (2022). Dominating risk impositions. *The Journal of Ethics*, *26*(4), 613–637.

Manson, K. (2023, July 5). *The US military is taking generative AI out for a spin*. Bloomberg.com. https://www.bloomberg.com/news/newsletters/2023-07-05/the-us-military-is-taking-generative-ai-out-for-a-spin

Martin, W., & Schinzinger, R. (1983). *Ethics in Engineering*. McGraw-Hill.

McDonald, J., Li, B., Frey, N., Tiwari, D., Gadepally, V., & Samsi, S. (2022). Great power, great responsibility: Recommendations for reducing energy for training language models. *arXiv Preprint arXiv:2205 09646*. https://doi.org/10.48550/arXiv.2205.09646

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*, *3*(1), 2053951716650211.

Ministerie van Algemene Zaken (2022, June 17). *Impact assessment fundamental rights and algorithms*. Report | Government.nl. https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, *31*(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*, *32*(2), 241–268. https://doi.org/10.1007/s11023-021-09577-4

Mollen, J. (2024). Towards a Research Ethics of Real-World Experimentation with Emerging Technology. *Journal of Responsible Technology*, 100098.

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, *3*(3), 869–877.

Nikolinakos, N. T. (2023). Ethical principles for trustworthy AI. *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related technologies-the AI act* (pp. 101–166). Springer International Publishing.

Pane, J., Francisca, R. D., Verhamme, K. M., Orozco, M., Viroux, H., Rebollo, I., & Sturkenboom, M. C. (2019). EU Postmarket surveillance plans for medical devices. *Pharmacoepidemiology and drug Safety*, *28*(9), 1155–1165. https://doi.org/10.1002/pds.4859

Peterson, M. B. (2013). New technologies should not be treated as social experiments. *Ethics Policy & Environment*, *16*(3), 349–351.

Polonioli, A., Ghioni, R., Greco, C., Juneja, P., Tagliabue, J., Watson, D., & Floridi, L. (2023). The Ethics of Online controlled experiments (A/B testing). *Minds and Machines*, *33*(4), 667–693.

Popper, K. (2013). *The poverty of historicism*. Routledge.

Porter, Z., Habli, I., McDermid, J., & Kaas, M. (2024). A principles-based ethics assurance argument pattern for AI and autonomous systems. *AI and Ethics*, *4*(2), 593–616.

Ranchordas, S. (2021). Experimental regulations for AI: sandboxes for morals and mores. *University of Groningen Faculty of Law Research Paper*, (7).

Raymond, N. (2019). Reboot ethical review for the age of big data. *Nature*, *568*(7752), 277–277.

Resseguier, A., & Ufert, F. (2024). AI research ethics is in its infancy: The EU's AI act can make it a grown-up. *Research Ethics*, *20*(2), 143–155.

Rheinberger, H. J. (1997). *Toward a history of epistemic things*. Synthesizing proteins in the test tube.

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and benefits of large language models for the environment. *Environmental Science & Technology*, *57*(9), 3464–3466. https://doi.org/10.1021/acs.est.3c01106

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large Language models a mirage? *arXiv Preprint arXiv:2304 15004*. https://doi.org/10.48550/arXiv.2304.15004

Seddon, P. (2023, September 29). *AI chatbots do work of civil servants in productivity trial*. BBC News. https://www.bbc.com/news/uk-politics-66810006

Smith, H. (2021). Clinical AI: Opacity, accountability, responsibility and liability. *Ai & Society*, *36*(2), 535–545. https://doi.org/10.1007/s00146-020-01019-6

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv Preprint arXiv:2206 04615*. https://doi.org/10.48550/arXiv.2206.04615

Stilgoe, J. (2016). Geoengineering as collective experimentation. *Science and Engineering Ethics*, *22*, 851–869. https://doi.org/10.1007/s11948-015-9646-0

Stilgoe, J. (2020). Who's driving innovation. *New Technologies and the Collaborative State. Cham, Switzerland: Palgrave Macmillan*.

Taylor, L. (2021). Exploitation as innovation: Research ethics and the governance of experimentation in the urban living lab. *Regional Studies*, *55*(12), 1902–1912.

Vaassen, B. (2022). AI, opacity, and personal autonomy. *Philosophy & Technology*, *35*(4), 88. https://doi.org/10.1007/s13347-022-00577-5

Van de Poel, I. (2013). Why new technologies should be conceived as social experiments. *Ethics Policy & Environment*, *16*(3), 352–355.

Van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics*, *22*(3), 667–686. https://doi.org/10.1007/s11948-015-9724-3

Van de Poel, I. (2017a). Society as a laboratory to experiment with new technologies. *Embedding new technologies into society: A regulatory, ethical and societal perspective*, 61-68.

Van de Poel, I. (2017b). Moral experimentation with new technology. *New perspectives on Technology in Society* (pp. 59–79). Routledge.

Van Heerden, A. C., Pozuelo, J. R., & Kohrt, B. A. (2023). Global mental health services and the impact of artificial intelligence–powered large language models. *JAMA Psychiatry*, *80*(7), 662–664. https://doi.org/10.1001/jamapsychiatry.2023.1253

Vitak, J., Proferes, N., Shilton, K., & Ashktorab, Z. (2017). Ethics regulation in social computing research: Examining the role of institutional review boards. *Journal of Empirical Research on Human Research Ethics*, *12*(5), 372–382.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., & Fedus, W. (2022). Emergent abilities of large language models. *arXiv Preprint arXiv:2206 07682*. https://doi.org/10.48550/arXiv.2206.07682

Wong, R. Y., Boyd, K., Metcalf, J., & Shilton, K. (2020, October). Beyond checklist approaches to ethics in design. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 511-517).

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in generative ai. *arXiv preprint arXiv:2403.02726*.