OJED
OPEN JOURNALS IN EDUCATION

Volume 16, Issue 2 (2024), pp. 86-94
Journal of Comparative & International Higher Education
DOI: 10.32674/jcihe.v16i2.5779 | https://ojed.org/jcihe

The Ethics of Research and Teaching in an Age of Big Data

David Lundie

University of Glasgow, UK

Corresponding author David Lundie: Email: david.lundie@glasgow.ac.uk
Address: University of Glasgow School of Interdisciplinary Studies, Crichton Campus, Dumfries, DG1 2JS, UK

This article was not written with the assistance of any Artificial Intelligence (AI) technology, including ChatGPT or other support technologies.

Abstract

Big Data offers opportunities and challenges in all aspects of human life. In relation to research ethics, Big Data represents a normative difference in degree rather than a difference in kind. Data are more messy, rapid, difficult to predict, and difficult to identify owners; but the principles of informed consent, confidentiality, and prevention of harm apply equally to digital data. Recognition that technologies are not inherently value neutral, and that data collection, aggregation, and their use in decision making can both create and intensify inequities and harms is central to applying these principles. Data justice extends concern with voice and authenticity into the digital domain. Universities act as gatekeepers to professional accreditation in fields including software engineering. The relation between academic freedom of enquiry, state and corporate interests in the Big Data age raises important questions about power and control in the academy, which have governance implications.

Keywords: assessment, big data, governance, large language models, research ethics

---

Introduction

Recent years have seen the exponential growth of Big Data analytics in many fields of human endeavor. Big Data has been defined as “high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation” (Gartner, 2015). Following this definition, not all large datasets will qualify as Big Data. Large standardized datasets (such as those produced by international Program for International Student Assessment’s educational assessments which produce static analyses) would not be included in this definition (Hartong, 2016). The uses of Big Data are very often removed from the kinds of processes commonly understood as data processing, such as conventional social research or educational assessment. Big Data can include data generated by such diverse areas as haptics (information on movement and non-conscious body activity such as heart rate and galvanic skin response), natural language, and environmental sensors in the

Received April 8, 2023; Revised August 1, 2023; Accepted September 1, 2023

Internet of Things (McEwen & Cassimally, 2013). These high-variety datasets require complex algorithms to analyze, the process of "automated reasoning" required is often quite distinct from the processes of abstraction, and hypothesis testing employed by human researchers (Reid, 2016). Further, data and information are not synonymous— information is well-formed data that are meaningful under some level of analysis (Floridi, 2004). The level of analysis can be as large as an entire city (Carta, 2019) or as intimate as the self (Sumartojo et al., 2016).

This essay seeks to understand the ethical challenges of Big Data for teaching, research and governance in global higher education. In this essay, I employ an ethical framework of rights, harms, and circumstances to understand the ways in which Big Data analytics in general, and large language models in particular, are operative in the academy. Highlighting challenges of clarity, transparency, and property rights in the datafied university, I argue that, although conventional ethical models can and should provide a guide to continued practice, these can become unduly complex and can obfuscate harms if attention is not paid to the structures and interests underpinning Big Data practices.

Digital Ethics

In an influential law paper in the early days of the commercially available Internet, United States appeals court judge Frank H. Easterbrook compared "digital law" to the "law of the horse." His point, principally, is that there are some cases in which the interaction of law intersect with some other fields, such as economics or international relations, which illuminates aspects of jurisprudence more broadly, but that computers, like horses, are not such a case (Easterbrook, 1996). While horses may at times appear in property law in relation to their ownership, or in relation to torts relating to damage done to property by horses, and at other times appear in the law as a mode of transport, regulated by the rules of the road, and still other times in relation to animal welfare law, a lawyer with any general knowledge of the law as a whole will be able to apply some common sense to understand which of these frames of reference apply to the given case. There is no need for a distinctive "law of the horse." Applying this general approach to digital law, Easterbrook argues for clear rules, transparent bargaining institutions, and clear property rights, enabling a liberal framework to operate in the digital field as it does in other fields of civil life.

Applying this approach to ethics today involves recognizing that the same ethical principles which operate in other spheres of life can be applied in the context of Big Data. In relation to the object of ethical action, persons continue to be imbued with intrinsic value, and the principle of avoiding deliberate harm and respecting individual autonomy continues to be relevant. These principles are agnostic to major debates in moral philosophy, such as between deontological and utilitarian ethics, and are similar to the broad consensus conditions of causal connection, knowledge of consequences and autonomy summarized by Noorman (2012). Morally salient circumstances can often be inferred from analogy to the physical world in relation to such matters as public and private, where gatekeeping structures operate in online spaces. In relation to the professional ethics of the academy, these public/private distinctions are usually quite well-defined: just as a teacher has a right to know what her students say about course content during a seminar but not in their dorm rooms after, so they have a right to access discussions in a university virtual learning environment, but not a private text chat; just as a researcher can treat letters to a newspaper editor as public documents but needs to gain ethical approval to survey members of a private club, so things posted on open web forums can be treated as public, but forums that require subjects to register or receive gatekeeper approval cannot. Thus far, there is no need for an "ethics of the horse."

In relation to transparency, clarity, and property rights, however, developments since Easterbrook's 1996 paper raise significant questions. The high velocity and variety of Big Data requires complex algorithms for analysis, and the use of heuristic machine learning algorithms often means that the complex multi-factor correlations they identify are not easily comprehensible to the human agents providing the data. Consider, for example, a hypothetical correlation between timing of cardiovascular activity and voter intent: it would not occur to a polling company to search for this correlation, nor would it occur to an individual purchasing a wearable heart monitor watch that their data might be used by polling companies for political advertising; nonetheless, Big Data often exploits such unexpected confluences in the data. This has clear implications for transparency and informed consent, as it points to the limits of our ability to conceptualize the uses to which our data may be put. In practice, many people pursue what Daniel Solove terms "security through obscurity"—believing their data is secure online because it would not be of interest to anyone (Hartzog & Stutzman, 2013). The ability to transparently understand the uses to which their data may be put is not merely a matter of reading the privacy statement, though one study put the opportunity cost of every user reading the privacy policy of every website they use at least once a year at 54 billion hours in the United States alone (this compares with 3.4 billion hours spent by every American taxpayer completing their income tax returns around the time of the study) (McDonald & Cranor, 2008). Rather, even if individuals were to consent to the uses to which their personal data were put by one website or another, it is in the sale and aggregation of this data and its potential secondary uses that transparency becomes near impossible.

Turning to clarity, a further use of Big Data by Large Language Models (LLM) is to construct artificial agents who can mimic natural language which is indistinguishable from a human language user. This has clear implications for the ways we understand the ethical harm principle. One of the earliest theoretical tests of Artificial Intelligence (AI), proposed by mathematician Alan Turing, was to posit that a machine is intelligent if its language use is indistinguishable from that of a human by a human interrogator (Turing, 1950). Subsequent critiques of this model have sought to highlight the difference between Turing's "parlor game" style test and a more expansive test of general intelligence, the need to attend to the subcognitive unconscious associative structures essential to human language use, and the link between cognitive and sensory information in human communication (French, 2000). As Artificial General Intelligence models relying on LLM approach Turing Test viability, a further significant question relates to the difference between "the strong results of reproductive, engineering AI, [relative to] the weak results of productive, cognitive AI" (Floridi, 2011). Reproductive AIs in this decade, such as ChatGPT, match patterns in existing human-created data—essentially constructing language through a complex version of "if the first N words of a sentence are this, human authors are most likely to place word x at N+1." These models are difficult to distinguish from human agents, but are incapable of producing any new thinking.

The implications of LLM reproductive AIs for our ethical principles of harm and intent are twofold. Firstly, as it becomes more difficult to distinguish human from artificial agents, there is a risk that we learn to ontologize ourselves heteronomously, in relation to non-human rather than human agents (Floridi, 2014). Interacting with one set of agents to whom we can do as we please because they are means to ends and incapable of suffering harm leaves us ill-prepared for interacting with other sets of agents who have intrinsic value and to whom we can relate, help, and hurt. Secondly, although artificial agents are not persons, the language datasets they draw upon reflect a totality of human experience. Large datasets which include biased human inputs can amplify that bias, as has been seen in predictive policing algorithms (Fountain, 2022; O'Donnell, 2019). The interaction between these two threats—interacting with ontologically "empty" agents and those agents reflecting and reproducing unethical human data, opens the possibility of a cumulative harm—human agents imitating the biases of the machines they have interacted with, and machines learning from and imitating the human data generated by those interactions.

Regarding the third of Easterbrook's criteria, in relation to property rights, data has become at once more and less a form of private property. As the forms of data and media of collection become more granular and varied, European legal discussions increasingly frame digital privacy as a matter of human dignity (Floridi, 2016), implying a primary, inviolable, human right rather than a secondary, instrumental property-type right to our data. At the same time, the commercial model developed by Big Data corporations, most notably by social media, treats the user not as customer but as product, selling increasingly fine-grained data for advertising purposes and manipulating affect to encourage engagement (Ghosh, 2020). This makes judging the private/public circumstances of any interaction more complex than the examples cited earlier in the paper. The interaction between the generation of student assessment data, its collection, aggregation and comparison by plagiarism detection software, and the asset generation model of the plagiarism detection software company, for example, is rarely explored. Further, the opportunity cost for not engaging in that data-for-service transaction is extremely high—a lack of credibility on the part of the academic institution, or the refusal of academic credit on the part of the individual student—raising important ethical questions in relation to informed consent.

In light of these significant complexities, then, is there a need for "ethics of the digital?" Is the field no longer analogous to Easterbrook's "law of the horse?" Returning to Easterbrook's criteria, this may be best understood as a transitional question. If the interaction of ethics with another field illuminates the ethical, then it is worthy of distinct study—such as in relation to politics, war, environment, inter alia. In relation to the digital, there are clear challenges which face society as a whole, yet there is not yet a clear understanding of the aims, ends, and purposes of these challenges. To take one field that has sought to conceptualize these challenges and threats, the Copenhagen School of security theory takes a multi-sectoral approach to understanding securitization. Each sector has its own internal motivating logic—the political sector is concerned with the institutions of politics and the binding idea of the state; the societal sector with the preservation of "we identities" and the economic sector with the profit motive and fiduciary responsibility, for example (Buzan et al., 1997). While recent iterations of security theory have attempted to conceptualize an informational sector, and enumerated information security threats, the definitions offered for informational security tend to focus on practices, such as communications and influence, digitization, methods and techniques of data transmission in networks (Ivancik, 2021), rather than identifying any sectoral logic to the informational sphere. Where motivations are attributed to information security actors, these either revert to the logic of the economic sector, or to actors aiming to secure or subvert political stability. Even turning to the technology ethics sector itself, a review of the leading Institute of Electrical and Electronics Engineers (IEEE) journal dealing with privacy ethics questions shows a disconnect between those papers engaging ethical questions and those papers reporting the engineering of technical solutions (Tse et al., 2015). In relation to digitalization, issues of law, security and ethics seem to be in agreement that there are aspects that illuminate the totality of the field, but also that we do not yet fully understand what aspects those are, what the implications are, and to what end. What of the educational field?

Digital Ethics in The Academy

Learning and Teaching

Given how deeply embedded Big Data is within our information management processes, it should not come as a surprise that learning and teaching practices are impacted. There is insufficient space here to address in full the intersection of epistemic virtue and data ethics, but a few specific examples can highlight the dangers of interpreting higher learning as though it were merely a process of information transfer (Lundie, 2016). In the European Higher Education Area, for example, the recipient of an undergraduate degree is expected to demonstrate knowledge and understanding in a field of study that is informed by knowledge of the forefront of the field, apply that knowledge through sustained argument, problem solving and critical analysis, and demonstrate the skills necessary to undertake further autonomous study (Bologna Follow-Up Group, 2005). These principles of autonomy, argument, and applied knowledge lend themselves to forms of assessment designed to measure originality and critical synthesis, not merely the transfer of information. Further, these principles draw on a long history of humanistic study in the European university—higher education is about the cultivation of educated persons, not simply knowledge acquisition, social reproduction, or technical competence.

In their present form, however, many of the forms of assessment employed by universities are dependent on LLMs and Big Data for their practical operation. This, in turn, leads to ongoing negotiation of questions of fairness and cheating. Most recently, the availability of open-access AI writing algorithms has raised concerns about students passing off algorithmically generated essays as their own work. Recalling that such LLM AIs are reproductive, matching patterns in existing human-created data, it is possible to see that the difference between software such as ChatGPT and a word-processor's built-in spelling and grammar checker are differences in degree, rather than differences in kind. Both operate by identifying the most likely sequences of words to appear in a positively received text. Yet in many cases, automated grammar checkers are encouraged, while AI writing apps are forbidden. Already, Turnitin (which relies upon an LLM which is constantly updated from essays submitted to its subscribers) has introduced features designed to detect AI-generated writing (Staton, 2023).

The potential harm to learner autonomy, and to knowledge itself, from these intersecting LLM systems, is rarely considered in relation to the architecture of the systems themselves. Reproductive AI relies on large language datasets, comprising all of the hitherto human-produced language content in a field. Detection systems rely on the scraping of AI-produced language in addition to human-produced language, as will future reproductive AIs. The result of this may be an increasingly narrow scope for human language to express originality, as our exposure to language becomes increasingly dependent on structures and patterns derived from the past. Merely trying harder to differentiate between human and artificial agents, as the cat-and-mouse generation/detection of informational content continues to consume one another's data, is not a viable solution.

More concerning than this particular development, however, are the ways in which attempts to exclude Big Data methods from our pedagogies and assessment practices have reshaped learning in unhelpful ways. Providing access to Turnitin scores in order to help students to self-diagnose poor scholarly practice, for example, can result in both an anxiety about the numerical score produced by the software, and a genuine confusion as to the relationship between language structure and originality. From experience, I have known students accused of misconduct incriminate themselves inadvertently by saying that they believed they had changed an idea enough to not be counted as plagiarism, believing this to be good practice. The availability of vast searchable academic databases such as Google Scholar can lend themselves to similar processes, whereby students first state an unsupported opinion, then find a scholarly source to back it up. These inadvertent, emergent biases in practice do not represent any devious intent on the part of the student, but rather reflect the confusion between reproductive, machine definitions of learning as information transfer, reverse-engineered from existing language data, and an authentically human definition of knowledge.

Possible solutions to these problems include a recognition that Big Data provides opportunities across the range of knowledge-intensive professions for which higher education provides a preparation. As in many other areas in which personal data is increasingly viewed as an aspect of the person, rather than as a property relation, this may involve more personal, enacted, less alienated forms of assessment, as well as raising student awareness of the threats and challenges of data-driven disinformation, including the passing off of AI-generated responses for those areas in which evidence of human autonomy is sought.

Research

With regard to research ethics, it is possible to follow the same principle of adhering to Easterbrook's "law of the horse" application of real-world principles to the digital up to the point that harms and rights attributions become too complex to disentangle clearly and transparently. This was the approach taken by the British Educational Research Association (BERA) in the reauthoring of its guidelines for ethical research in 2018. In relation to informed consent, for example, the guidelines advise researchers as follows:

> Where research draws on social media and online communities, it is important to remember that digital information is generated by individuals. Researchers should not assume that the name given and/or identity presented by participants in online fora or media is a "real" name: it might be an avatar. This avatar could represent a human or a bot, but behind either will be one or more human creators responsible for it, who could therefore be regarded as participants; whether and how these potential participants might be traceable should be considered. Where an organization shares its data with researchers, those researchers have a responsibility to account for how and with what consent that data was gathered; they must also consider the authorship of that data and, consequently, whether it is necessary to independently approach the relevant individuals for consent concerning its use. Researchers should keep up to date with changes in data use regulations and advice. (BERA, 2019, p. 7)

To highlight two key points in this paragraph: firstly, in relation to given identities, the guidelines highlight the importance of considering the human individuals behind the creation of digital personae. While a digital avatar may constitute a form of performance (Papacharissi, 2012), sometimes curated by a number of individuals on behalf of a high-profile individual, the new problem posed by LLMs is that the individual human creators who provided the language reconstructed by the algorithms are so far removed from the responses that it becomes impossible to attribute ownership rights over the text. Even if such attribution were possible, the number of creators involved would make any attempt to independently approach them for consent prohibitive.

Secondly, the guidance considers cases where organizational ownership is asserted over data. In almost every sphere of social life, organizations hold data on service users for a range of purposes. Within education, this can include some of the measures and metrics identified in the foregoing section. The level of consent given to the collecting organization is often tacit, implied, or given under some measure of duress—universities have always collected and collated data on student assessment performance, and Big Data does not introduce any novel data collection harms, but it does potentially change the data processing and aggregation climate in important ways. There are challenges of informed consent in organizational contexts that involves recognizing so that data may be shared up long and complex hierarchies. This includes sharing from individuals to academics grading their work, from those academics to university administrators seeking to understand patterns in departmental performance, from universities to Big Data corporations, perhaps contracted by national governments to carry out evaluations of the higher education sector as a whole, but who nonetheless reinscribe that data in line with their own collection processes, as highlighted earlier. Depending upon which level researchers seek to engage, the organizational data they collect may have undergone multiple mutations of consent.

The legal scholar Daniel Solove suggests a taxonomy of 16 distinct privacy harms, relating to four domains of information collection, processing, dissemination, and invasion (Mulligan, et al., 2016; Solove, 2008). This taxonomy suggests that privacy needs to be understood not as a single thing but a collection of related concepts with a family resemblance between them. We can use these four domains to understand the research process. With regard to invasion, the rules of informed consent as operative in real-world empirical research are relatively clear-cut—a researcher seeking direct access to a participant's private life needs to seek consent. In relation to the other three domains, however, research with Big Data is more complex. To return to the assessment data example, information collection may not have changed significantly from the time of pen-and-paper examinations, but the processing of this data for a multitude of purposes other than the assigning of a grade, and its dissemination to other mediating organizations, has changed significantly. While the values-in-design literature contains a number of practical suggestions for embedding privacy-protection strategies in the design of data systems (Flanagan et al., 2009; Wicker & Schrader, 2010), researchers are rarely in the position of designing

Big Data systems for the collection and processing of data, but rather of being secondary users of that data. The BERA guidelines continue:

> Anonymity is much harder to guarantee in digital contexts. The policies of some social media sites which require identification at signup may exacerbate this. Researchers need to be aware that participants' understandings of their level of privacy in a particular online space may be inaccurate. Ambiguity about privacy within some online communities in which sensitive or illegal topics are being discussed, or material shared, raise(s)further ethical concerns. Relatedly, researchers should consider the question of what online content, in what circumstances, they would be obligated to report to relevant authorities and/or online service providers, bearing in mind any agreements entered into regarding confidentiality and anonymity… Researchers using data gathered in such contexts should inform the community concerned about how the data will be used. (BERA, 2019, p. 23)

A further point, not recognized in the BERA guidelines, is that researchers engaging in data collection on social media need to be aware that they are inside the algorithm ecosystem they seek to research. Two solutions present themselves to this problem. The first involves a further recourse to Big Data algorithms to analyze the large language datasets generated by Application Programming Interfaces (APIs), such as Twitter's real-time streaming API, or even Twitter's "firehose," which provides access to all 400m daily tweets on the platform. Such analytics require a unique skillset for social researchers. The other solution is for researchers to recognize their own digital positionality. A normal Twitter keyword search, for example, will not return all of the items which include a particular keyword, but a curated sample filtered by the algorithm, filtered based on past engagement patterns of the user searching for them. The creation of avatars accounts to understand an alternate positionality—for example, exploring what a social media platform presents about the reliability of health journalism to a user who follows and likes US Republican media content, and how that varies from the search items displayed to a user who follows and likes Canadian Liberal political content—presents further ethical questions around misrepresentation. Such an approach may technically constitute a deception study; however, it is unclear who, if anyone, is being deceived. Failing to recognize this positionality undermines the rigor and reliability of much small-scale qualitative social media research.

In practice, these changes lend themselves to two changes in the ways social researchers do their work. The first relates to the proliferation of publicly available and proprietary datasets which can provide whole-population data on a range of themes. In practice, the levels of statistical understanding and the computing power necessary to carry out complex multivariate work with these datasets has tended to make these the preserve of specialist units, such as the University of Glasgow's Urban Big Data Centre. This kind of large-scale, government funded approach to urban analytics has potential to greatly improve wellbeing, but also raises significant questions, not only regarding the consent of participants involved in the datasets, but also regarding the scale of governance. Such services are expensive, and tend to be at the service of governments or multinational corporations. The Urban Big Data Centre has been pioneering ethical approaches to participatory coproduction research with the end users of such services. One example is the Waterproofing Data project (Pitidis et al., 2022) which sought to involve young people affected by climate change in the Global South in developing solutions to the problem and understanding and negotiating the data processing challenges posed by those solutions.

The second change relates to the blurring of the boundary between secondary and primary data. As the publication costs of publicly available sources tend toward zero (Weinberger, 2011), many of the private opinions, professional judgments, management decisions, and mid-level policy recontextualizations which would previously have remained private, circulated as memos within an office building, or only available through oral interactions, have become accessible from without. As noted above, however, this is not the same as this data becoming "publicly available" in the sense that is ethically pertinent. This can mean that it is unclear when social research leaves the "literature" phase and becomes empirical. At the University of Glasgow, for example, empirical research usually requires the completion of an ethical approval form, together with participant information letters and consent forms. This is reviewed by a College Ethics Committee, under conditions set out in the UK Concordat to support research integrity (UKRIO, 2019). Research involving social media, Big Data, or other sources in this blurred middle, however, requires a different ethical approval form for research with "non-standard data." Although this is considered by the same committee process, it does suggest that the considerations for an ethics of the digital are at least procedurally distinct from those of common research ethics. This process tends to foreground the complexity and indeterminacy of big data research, providing participants with a reading list of potentially pertinent journal articles, professional guidelines, and methodological guides (University of Glasgow, 2023).

Governance and Administration

Through their research and teaching function, universities play an important role as gatekeepers to professional employment. For this reason, concerns that the technology sector is cannibalizing the governance of higher education (Lundie et al., 2022) ought to be taken seriously. The corporatization of higher education, particularly in the digital sector, entails that the burden of responsibility for preparing talent for the challenges of work appears to be fundamentally shifting; research and teaching are "outsourced" to the university, but metrics of quality increasingly are set and adjudicated by graduate employment in high status corporations. New data technologies have been identified as both an effect of these international corporatization processes and a driving force in their governance (Hartong, 2016).

When heads of government and chief executives of Big Data corporations met to discuss the future of education in the world economic forum in Davos in January 2020, they agreed that the education sector is due for an overhaul to make schools and universities fit for the fourth industrial revolution. This "Education 4.0" model would have to align skills to fit the needs of the corporate sector (World Economic Forum, 2020).

While the history of universities predates state involvement in education, and always conferred measures of academic freedom over and against the authority of the state, since the industrial and democratic revolutions of the 19th century, education governance has been seen as a prerogative of state sovereignty. The principle of academic freedom comprises freedom of inquiry in research and the freedom to teach or communicate ideas and facts. Against the backdrop of technical and corporate takeover, some of the reactionary exercises of state authority to limit, for example, the interpretation of history and politics (Miller et al., 2023; Woolcock & Zeffman, 2017) can be understood as a rearguard action in a climate in which other forms of interference in curriculum, through quasi-markets of educational goods reinscribed as currency according to corporate metrics (Lundie, 2022), have become normalized. The differences between universities and the research labs of the tech sector begin to blur, with new research ideas that inform the tech industry coming equally from publicly funded grants to university research labs and the private labs of industry, and researchers moving seamlessly between university and corporate roles.

To marshal these competing state and industry imperatives, international organizations increasingly play a role in reifying and standardizing measures of educational effectiveness, which in turn drive governance policies of leading universities. Given the impact universities can have on the economic attractiveness and investment potential of nations, these governance demands in turn drive education policies across the world. The incursion of technological algorithms, themselves the proprietary secrets of private providers, in this process is an ethical blind spot in current thinking. From international rankings (e.g., Quacquarelli-Symonds; Academic Ranking of World Universities; and Times Higher Education World Rankings) to the role of large educational conglomerates such as College Board and Pearson Educational, such institutions play important gate-keeping roles in selecting which universities' research is funded and achieves impact, and selecting which students can access university qualifications that prepare them for high status professions. An example of the influence of these ranking systems on the prerogatives of state sovereignty is provided by the UK High Potential Individual visa scheme. This visa scheme offers the opportunity to live and work in the UK to graduates of the global top 50 universities, defined as an institution that has appeared in two of the three global ranking systems listed above (Nietzel, 2022), essentially ceding control of its borders to proprietary corporate algorithms.

Implications and Conclusion

From Easterbrook's concern with clear rules, transparent bargaining, and clear property rights in the digital sphere, it has been possible to theorize three constellations of normative Big Data questions operative in higher education today. Firstly, in relation to transparency, the risks of harm in relation to learners ontologizing themselves and their knowledge in relation to artificial agents has been explored in relation to large language model reproductive AI. Secondly, in relation to transparent bargaining, the challenges to recognizing and respecting the intrinsic value and autonomy of human research subjects when datasets become infinitely reproducible, subject to portability and recombination across domains, and the unexpected results that can arise from the automated reasoning processes involved in Big Data analysis pose difficult questions for informed consent in research. Finally, in relation to property rights, attention needs to be drawn toward the impact of corporate interests on freedom and consent in higher education governance processes. In all three cases, awareness of the threats and opportunities posed by Big Data, and the structures and algorithms which generate these, are necessary to enable students, researchers, and university administrators to contextualize and navigate these ethical dilemmas, maintain clarity on their value for the human actors within the system, maximize benefits, and minimize harms.

Ethical theory at present stands at a transitional point, not yet having arrived at a distinctive *telos* and institutional logic of the technological sector that would illuminate a distinctive "ethics of the digital," and yet finding it increasingly difficult to proceed without one. As harm and intent become more remote from the human causal agent, traditional utilitarian and deontic ethical calculations become more difficult to apply. These issues concern technologies whose influence is felt globally, requiring nuanced and rapid response, yet the resources to address them remain concentrated in the Global North, potentially exacerbating inequalities for universities in the Global South. As data becomes more than merely a piece of personal property, indeed, in much current research the most fundamental properties of the human person such as genetic structure and neural activity become datafied, approaches to data management that are grounded in property relations become insufficient, not only in research ethics but in the wider world. Infinite reproducibility of data and language holds out a challenge and also a promise to educators, offering the potential to upskill graduates in literate domains in ways analogous to the impact the introduction of spreadsheets and calculators had on mathematical domains—freed from the labor-intensiveness of calculating complex statistical significance tests manually, it becomes possible to advance more quickly to higher level analytical skills, for example. The same challenge calls us to a more profound ethical engagement with the infinite reproducibility of data generated in the course of research and evaluation, its appropriation by corporate technological interests, and infinite manipulability by hitherto uninvented large language machine learning models, and the consequences of this for a still more accelerated and undifferentiated world.

References

British Educational Research Association. (2019). Ethical guidelines for educational research, fourth edition (2018). https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018-online

Bologna Follow-Up Group. (2005). An overarching framework of qualifications for the EHEA. http://www.ehea.info/cid102059/wg-frameworks-qualification-2003-2005.html

Buzan, B., Waever, O., & de Wilde, J. (1997). Security: A new framework for analysis. Lynne Rienner Publishers.

Carta, S. (2019). Big data, code, and the discrete city: Shaping public realms. Routledge. https://doi.org/10.4324/9781351007405

Easterbrook, F. H. (1996). Cyberspace and the law of the horse. *University of Chicago Legal Forum*, 207–216. https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=2147

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2009). Embodying values in technology: Theory and practice. In *Information technology and moral philosophy.* Edited by J. Van Den Hoven, & J. Weckert, (pp. 322–353). Cambridge University Press. https://doi.org/10.1017/CBO9780511498725.017

Floridi, L. (2004). Information. In *The Blackwell guide to the philosophy of computing and information.* Edited by L. Floridi, (pp. 40–61). Blackwell. https://doi.org/10.1002/9780470757017

Floridi, L. (2011). Children of the fourth revolution. *Philosophy & Technology*, 24, 227–232. https://doi.org/10.1007/s13347-011-0042-7

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press. https://www.oii.ox.ac.uk/research/publications/the-fourth-revolution/

Floridi, L. (2016). On human dignity as a foundation for the right to privacy. *Philosophy & Technology*, 29, 307–312. https://doi.org/10.1007/s13347-016-0220-8

Fountain, J. E. (2022). The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. *Government Information Quarterly*, 39(2), 101645. https://doi.org/10.1016/j.giq.2021.101645

French, R. M. (2000). The Turing Test: The first 50 years. *Trends in Cognitive Science*, 4(3), 115–122. https://doi.org/10.1016/S1364-6613(00)01453-4

Gartner. (2015). Information technology glossary. https://www.gartner.com/en/information-technology/glossary/big-data

Ghosh, D. (2020). Terms of disservice: How Silicon Valley is destructive by design. Brookings Institution Press.

Hartong, S. (2016). Between assessments, digital technologies and big data: The growing influence of 'hidden' data mediators in education. *European Educational Research Journal*, 15(55) 523–536. https://doi.org/10.1177/1474904116648966

Hartzog, W., & Stutzman, F. (2013). The case for online obscurity. *California Law Review*, 101, 1–50. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1597745

Ivančík, R. (2021). Security theory: Security as a multidimensional phenomenon. *Vojenske Reflexie*, 16(3), 32–53. https://doi.org/10.52651/vr.a.2021.3.32-53

Lundie, D. (2016). Authority, autonomy and automation: The irreducibility of pedagogy to information transactions. *Studies in Philosophy and Education,* 35, 279–291. https://doi.org/10.1007/s11217-016-9517-4

Lundie, D. (2022). *School leadership between community and the state: The changing civic role of schooling*. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-99834-9

Lundie, D., Zwitter, A., & Ghosh, D. (2022, January 31). Corporatized education and state sovereignty. https://www.brookings.edu/blog/techtank/2022/01/31/corporatized-education-and-state-sovereignty/

McDonald, A. M., & Cranor, L. F. (2008). The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4(3), 543–568. https://www.technologylawdispatch.com/wp-content/uploads/sites/26/2013/02/Cranor_Formatted_Final1.pdf

McEwen, A., & Cassimally, H. (2013). *Designing the internet of things*. John Wiley & Sons. https://www.wiley.com/en-us/Designing+the+Internet+of+Things-p-9781118430620

Miller, V., Fernandez, F., & Hutchins, N. H. (2023). The race to ban race: Legal and critical arguments against state legislation to ban critical race theory in higher education. *Missouri Law Review.* 88(1), 1–46. https://scholarship.law.missouri.edu/mlr/vol88/iss1/6/

Mulligan, D. K., Koopman, C., & Doty, N. (2016). Privacy is an essentially contested concept: A multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A*, 374(2083), 1–17. https://doi.org/10.1098/rsta.2016.0118

Nietzel, M. T. (2022, May 31). Britain opens up its visas for graduates of world's top universities. https://www.forbes.com/sites/michaeltnietzel/2022/05/31/britain-opens-up-its-visas-for-graduates-of-worlds-top-universities/?sh=464e8a827fcf

Noorman, M. (2012). Computing and moral responsibility. http://plato.stanford.edu/archives/fall2012/entries/computing-responsibility

O'Donnell, R. M. (2019). Challenging racist predictive policing algorithms under the equal protection clause. *New York University Law Review*, 94(3), 544–580. https://www.nyulawreview.org/wp-content/uploads/2019/06/NYULawReview-94-3-ODonnell.pdf

Papacharissi, Z. (2012). Without you, I'm nothing: Performances of the self on Twitter. *International Journal of Communication*, 6, 1989–2006. https://ijoc.org/index.php/ijoc/article/view/1484/775

Pitidis, V., de Albuquerque, J. P., Coaffee, J., & Lima-Silva, F. (2022). Enhancing Community Resilience through Dialogical Participatory Mapping. In *ISCRAM* (pp. 495-503). https://www.idl.iscram.org/files/vangelispitidis/2022/2435_VangelisPitidis_etal2022.pdf

Reid, D. (2016). Man vs. machine: The battle for the soul of data science. In *Big Data Challenges: Society, Security, Innovation and Ethics*. Edited by A. Bunnik, A. Cawley, M. Mulqueen, & A. Zwitter, (pp. 11–22). Palgrave. https://doi.org/10.1057/978-1-349-94885-7_2

Solove, D. J. (2008). *Understanding privacy*. Harvard University Press. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1127888

Staton, B. (2023, April 3). Universities express doubt over tool to detect AI-generated plagiarism. https://www.ft.com/content/d872d65d-dfd0-40b3-8db9-a17fea20c60c

Sumartojo, S., Pink, S., Lupton, D., & Heyes LaBond, C. (2016). The affective intensities of datafied space. *Emotion, Space and Society*, 33–40. https://doi.org/10.1016/j.emospa.2016.10.004

Tse, J., Schrader, D. E., Ghosh, D., Liao, T., & Lundie, D. (2015). A bibliometric analysis of privacy and ethics in IEEE Security and Privacy. *Ethics and Information Technology,* 17, 153-163. https://doi.org/10.1007/s10676-015-9369-6

Turing, A. M. (1950). Computing machinery and intelligence. *Mind,* 59(236), 433–460. http://www.jstor.org/stable/2251299

UK Research Integrity Office (2019). *Concordat to Support Research Integrity*. https://ukrio.org/research-integrity/the-concordat-to-support-research-integrity/

University of Glasgow (2023). *Online Information Links for Internet Based Research*. https://www.gla.ac.uk/colleges/socialsciences/students/ethics/ethicstrainingresources/onlinedatainformationlinks/

Weinberger, D. (2011). *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room*. Basic Books.

Wicker, S. B., & Schrader, D. E. (2010). Privacy-aware design principles for information networks. *Proceedings of the IEEE*, 99(2), 330–350. https://doi.org/10.1109/JPROC.2010.2073670

Woolcock, N., Zeffman, H., & Geddes, D. (2017, Oct 25). Tory whip 'wanted names of Brexit lecturers for book research.' https://www.thetimes.co.uk/article/i-want-names-of-brexit-lecturers-tory-whip-chris-heaton-harris-tells-universities-6sv98nn0x

World Economic Forum. (2020). Education 4.0. https://initiatives.weforum.org/reskilling-revolution/education-4-0

DAVID LUNDIE, PhD. Senior Lecturer in Education and Deputy Head of School, University of Glasgow, UK. Principal Investigator: Teaching for Digital Citizenship: Data Ethics in the Classroom and Beyond. Deputy Editor: *British Journal of Religious Education*. david.lundie@glasgow.ac.uk