Author(s): Walter Barta

The "Great Unread" and the "Black Box"

Abstract:

Imagine that you can summarize and analyze the "Great Unread", all the works of literature in human history, but only by feeding it through a "Black Box", an algorithm that nobody fully understands. Does the output have explanatory value? If so, what kind?

Agenda:

Author(s):

Walter Barta:

- Digital Research Commons, M.D. Anderson Library, University of Houston, 4333 University Drive, Houston, TX 77204-2000
- ✉ wjbarta2@uh.edu, 💻 https://libraries.uh.edu/drc/

ISSN 1614-1687

Walter Barta is a principal investigator at the Digital Research Commons at the M.D. Anderson Library at University of Houston. He is also a recurring contributor to the Popular Culture & Philosophy book series with a special interest in science fiction and philosophy.

Case Description

Imagine that, as has been happening now for years, the digital humanities lab in the library at your university introduces a brand new digital tool that uses machine learning techniques to analyse its archives; the tool is incredibly useful to the extent that it can rapidly analyse all of the texts that humans have not had the time to read and produce interesting results: summaries of and comparisons between large groups of texts. These results have substantial implications for studies of the history of literature and comparative literature. But the tool is also highly suspect because nobody really knows how it works. The administration of the university encourages the use of the tool because they believe they can achieve rapid and intense publicity from its usage—not to mention it was expensive to develop. The faculty are generally impressed by (if a bit sceptical of) the results that the tool produces. Even the hardcore naysayers feel the pressures of "publish or perish" and acknowledge how the tool might assist them in meeting their deadlines. Humanists in general are enthusiastic that the tool can provide insights about the "great unread" works of literature accumulated through history (Cohen in Moretti, 2000). However, when asked, the computer scientists who developed the tool fully (if a bit embarrassedly) admit that the device is a "black box" that even they do not currently understand—and may never completely understand (Nature, 2023).

Questions

1. In what cases is it reasonable or expected for a digital humanities researcher to publish conclusions that were derived from digital methodologies that they do not (and perhaps cannot) understand?

2. In what cases is it reasonable or expected for a digital humanities researcher to publish conclusions that are derived from simple, explicable programs that they do understand but that are not state-of-the-art?

3. Is the trade-off between insightfulness and explicability better satisfied by finding a middling compromise? Or by finding minimal acceptable thresholds?

4. Do the scopes and methods of digital humanities research warrant an entirely new publication paradigm?

5. Is it really knowledge if we cannot explain it? Are the results still valuable even if unexplainable? To what extent is "Because the computer said so!" a reasonable explanation?

6. How could the humanists know that the results of a black box were interesting? How could they test the results? Would testing make it acceptable to publish the results?

7. Can the output of a black box provide insight if you do not understand how it works? Should researchers trust the research and tools of others?

8. Does anyone understand the stack of technologies we call a computer well enough to be confident of any computational results? If the computer itself is a black box to most researchers, then why do we trust it in academic work?

9. Should the humanists have trusted a tool developed from another discipline? Would it have made a difference if the tool had been programmed by digital humanists?

10. If there was an error in the results about a particular book, who would be responsible? The computer scientists who developed the tool? The lab that set it up and made sometimes erroneous results available? The administration who pressured the lab to publicize the project?

Exercises

1. Imagine you are running a small digital humanities laboratory. What methodologies would you employ? What is the most inexplicable tool you are willing to use? What is the least advanced tool you are willing to use?

2. Imagine you are tasked with designing a digital humanities research repository. How would you design it? What standards would you use? How would you categorize the projects?

Principle	Application (If Any)
Nonmaleficence	
Beneficence	
Respect for Autonomy	
Justice	
Explicability	
Accountability	

Normative Theories

Apply the different normative theories explained in the primer at the beginning of this volume to bring out issues in the case that might have been overlooked. How would a utilitarian approach this case? A deontologist? A virtue ethicist? Do any of these approaches accord with your own moral judgments or not? Do they arrive at the same verdict as the principles of AI ethics?

Expert Analysis (Read After Doing Your Own Analysis!)

In academic institutions around the world, digital humanities programs have been established with the intention of bringing computational techniques and quantitative methodologies to humanities subjects (Presner, 2009). These developments have been part of a century spanning project to bridge the divide between the "two cultures" of academia, the natural sciences, and the liberal arts, in an attempt to bring humanity to the former and bring methodological rigor to the latter (Snow, 1959). To this end, over the course of decades, various datamining techniques, ranging from basic word frequency searches to statistical analyses, been employed on

massive bodies of literature to discover new insights; while, simultaneously, scholars have attempted to apply their critical, humanistic lenses to these emerging technologies.

There are several grand humanistic benefits to the digital humanities research agenda. As Franco Moretti and Margaret Cohen point out, digital processing allows us to discover the long-lost treasures of textual history by examining the depths of the "great unread," all the ubiquitous but anonymous works of literature that failed to catch on culturally and commercially (Cohen in Moretti, 2000). As Ted Underwood has put it, the digital humanities allow researchers in many fields to explore "distant horizons", absorbing entire libraries all at once and drawing big-picture conclusions from the vastness of the data that never could have been comprehended without the memory and speed of computers (Underwood, 2019).

However, with the development of more sophisticated computational techniques, the capabilities of digital humanities research are changing faster than many scholars can keep up with. Furthermore, many of these digital models are by design "black boxes," complicated systems with inscrutable contents, which means that how they work is inexplicable most if not all humanists—and even to the computer scientists (Nature, 2023). This inexplicable property makes these complex tools increasingly awkward components in academic research, in contrast to the simple, easily understood search tools of decades past.

So, drawing our attention to the ethics of using such models, we may follow W. D. Ross by considering a plurality of prima facie (first glance) goods, including justice, virtue, pleasure, and knowledge (Ross, 1930). We might analyse the results of a given research methodology based on the production of these Rossian goods. Of these goods, we may primarily narrow our focus onto knowledge as an intrinsic good, since the pursuit of knowledge is a special goal for academic institutions, and therefore the dominant good towards which academic research projects, including digital humanities projects, are directed. For our purposes, without getting bogged down by too much epistemology, we can accept a working definition of knowledge in accordance with the ancient definition that Plato plays with: "justified, true belief" (Chappell, 2023). Thus, a good digital humanities research project will produce knowledge by verifying and justifying beliefs about humanities topics using digital techniques. In contrast, a bad research project will produce no knowledge, or (worse still) will produce ignorance by producing justified false beliefs or dogma by producing unjustified true beliefs. We may further consider that knowledge can also have extrinsic usefulness in addition to its own intrinsic value. And, in addition, we may consider the good of the virtues of the researchers as researchers as knowledge-gatherers, chiefly the epistemic virtues of curiosity and knowledgeability.

When considered in terms of the intrinsic and extrinsic value of knowledge, the question boils down to a relatively simple set of questions about benefits and costs (not unlike a utilitarian analysis). How much direct benefit in the form of knowledge does the black box methodology produce? Furthermore, how much indirect benefit is produced as a knock-on effect of that knowledge? And how much does the method cost, directly and indirectly? To the extent that the ends justify the means, as long the black box methodology produces knowledge without overriding costliness, the methodology can be reasonably justified.

When considered in terms of the virtue of the curiosity and knowledgeability of the researchers, the question is similar. How much epistemic virtue in the form of curiosity and knowledgeability do the researchers exercise by carrying out the project? How much vice do the researchers exercise? As long as the conduct of the project is discernibly an exercise in epistemic virtue, we can call it good.

Indeed, in many cases, we can easily imagine the above two conditions (knowledge and virtue) met, especially if the effectiveness of the methodology is great. The universities may get their funding. The faculties may get their tenure. Civilization may get more books for the archive. But, most importantly, the goodness of knowledge is produced and knowledgeability is exercised.

However, there are technical reasons to worry that knowledge and virtue may be conceptually impossible (or at least exceedingly difficult) in specific black box cases.

First, regarding the knowledge-products, as long as a research methodology involves a black box, we may have some doubt that knowledge can be produced at all. If, in a given instance, the researchers do not understand

how a method arrives at its conclusions, then the method may be unjustified in that instance; and if, in principle, the researchers cannot understand, then the method may be unjustifiable in general. If true, this would mean that black boxes do not or cannot produce knowledge. The problem is twofold: on the one hand, the black box may be producing false beliefs, and the researchers may not understand enough to reject them; on the other hand, the black box may be producing true beliefs that nonetheless have no understandable justification. In this sense, the black box case bears some resemblance to a so-called "Gettier Case," a case in which random chance nontrivially affects the justification of knowledge (Gettier, 1963). Like a stopped clock may be correct twice per day, a black box model may be correct sometimes but may be incorrect other times, for reasons that may be entirely contingent and independent of justificatory procedure. Thus, every proposed research question using such a methodology has a dilemma between the Scylla of dogmas (unjustified true beliefs) and the Charybdis of self-deceptions (justified false beliefs). This ironically leaves the humanists in a similar position as where they started, trading the inexplicability of the "great unread" with the inexplicability of a "black box"—the unknown you know for the unknown you do not.

Second, regarding the virtues of the researchers, such black box research cases are strange because in the limit case, where the black box does everything, the entire research process can be conducted without any reference to the researcher's own knowledgeability. To the extent that the research procedure is internal to the black box itself, the researcher has in a sense not participated in any critical step of the knowledge production line. Conceivably, a completely uncomprehending button-pusher could run the digital tool and leave the human expert out of the procedure entirely. This is a strange situation because it means that knowledge could be produced without anyone knowing it.

Perhaps we might counter the above issues by pointing out that most of what we believe is not known to us with scientific rigor anyway. Most of us have not ourselves conducted the justificatory research underlying our beliefs; instead, we refer to "the experts," "the literature," or "the field," and that is justification enough, practically speaking. So why can't we refer to a black box as our justification? In a sense we can: rather than making an "appeal to authority" we can make an "appeal to the box." And this is not even a problem for our casual, everyday needs. After all, we do not need to know how complex machines (cars/microwaves/computers) work to use them and reap their benefits. However, while perhaps not a problem for people generally, this is more of a problem for knowledge gatherers particularly, since understanding results may depend upon understanding methods. The "appeal to the box" only gives us conditional beliefs (conditional upon the reliability of the box) and not understanding, which is inadequate if we are the experts whose role it is to understand. It is one thing to say that laypeople have unjustified beliefs; it is quite another thing to say the "experts" do.

This picture is complicated further by the fact that most human knowledge gathering is done via complicated technological apparatus. Indeed, arguably, humans rarely if ever think unassisted; we use a variety of means to augment observation and cognition, such that our epistemology is much more distributed throughout a sociotechnical system than centralized in the human mind (Sloman & Fernbach 2017). To use the simple example of Gettier's clock, we do not even demand ourselves to know exactly how a clock works when we tell the time. The technological dependency just scales up when we consider state of the art science. Physicists use the Large Hadron Collider (the biggest particle accelerator on earth) to discover new subatomic particles, and yet it is unlikely that any single human physicist knows exactly how such a complicated machine works. Nonetheless, whether physicists use clocks or particle accelerators, we still largely accept their results as a form of knowledge.

One might argue that the black box is fundamentally different from these other examples because, whereas a clock and a particle accelerator can in principle be studied and understood with enough time and effort, the black box perhaps cannot be understood even in principle. This is perhaps an open question: AI experts may someday have efficient enough methods to trace back all of the trillions of parameters in the black box and explain them, though these methods may still be practically infeasible because, like some decryption algorithms, they could take longer than a human lifetime to run.

Either way, we are still faced with the challenge that unless every knowledge-gatherer fully understands every device they use, knowledge-gathering roles will depend on trusting others work without understanding it. So,

it is not necessarily that a failure to understand a methodology makes all beliefs in the results unjustified; rather many of our justifications must presuppose the reliability of the methodology to even get off the ground. What contexts allow for justifications of this kind is perhaps a question for further epistemological consideration.

Perhaps with the development of black box technologies we are entering a new scientific paradigm. As Moretti and Underwood themselves admit, no one can read a million novels. So perhaps the more realistic research question becomes less about understanding, via interpretation, than about explanation, via observation. So, perhaps while distant reading cannot say anything about individual novels, it can say a good deal about the history of novels. Many other disciplines use the statistical techniques to show correlations even when unable to demonstrate causation, and maybe literary theory is just in the early stages of this paradigm shift.

So, do these considerations rule out the use of black boxes? Certainly not. The methodological trade-offs described for black boxes may persist across the entire spectrum of digital tools, such that retreating to older methods may simply be retreating to older difficulties. On the one hand, the less simple the computational tool, the less explainable the results, but the more complex the tool, the more comprehensive the results; on the other hand, the more simple the computational tool, the more explainable the results, but the less complex the tool, the less comprehensive the results. In other words, there may be no perfect methods; there may always be a methodological trade-off between our ability to have compelling conclusions and our ability to have compelling explanations. Knowledge remains difficult to discover—perhaps until we fully illuminate the black boxes.

In the meantime, one straightforward way to render the black box dilemma more tractable would be to establish simple standards of transparency. Most disciplines now have rules about using opaque systems and require some levels of transparency. For instance, the Artificial Intelligence and Data Act prescribes transparency as one of the criteria for usage in high-impact domains (AIDA, 2023). This requirement of transparency is part of the growing global consensus around the best practices for operating such devices. The standard of transparency has at least two elements, 1) disclosing usage of the tool, and 2) verifying outputs of the tool. For instance, in medical contexts, diagnosis and treatment-planning by means of black box is generally not accepted. In such contexts, black boxes are used more to supplement than replace human decision-making, and as such disclosure of usage and verification of information are always necessary follow-up steps. As understanding of the technology further develops, a third element, 3) understanding the inner workings of the device, will increasingly become another expected standard of usage, especially for high impact use-cases. In the case of our black box, although the academic use-cases may be low-impact relative to comparable medical use-cases, the impact overtime of the erosion of academic standards may be high. So, implementing standards of transparency would still have importance. If the scholars reported using the device and double-checked its outputs before publication then the results could be considered significantly less questionable than otherwise, increasing their academic credibility.

In conclusion, through considerations of knowledge and virtue, one can observe that digital research technologies offer trade-offs between epistemological values: truths versus justifications (conclusions versus explanations). Whether these trade-offs will persist with future iterations of digital tools is yet to be seen, but we can suspect familiar institutional pressures to persist in the meantime, placing academics in the position of some potentially hard methodological compromises. How things will develop, the box only knows.

Reflection

Did you touch on everything this expert analysis identifies in your own analysis of the case? What did you miss?

Did you think of anything that could be added to the analysis or were there any points of disagreement? Be sure to justify your response with reasons.

References

Chappell, Sophie-Grace, "Plato on Knowledge in the Theaetetus", The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/fall2023/entries/plato-theaetetus/.

"ChatGPT is a black box: how AI research can break it open." Nature, 619, 671-672 (2023). doi: https://doi.org/10.1038/d41586-023-02366-2.

Presner, Todd et al. (2009). "The Promise of the Digital Humanities." Humanities Blast, https://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf.

Gettier, Edmund (1963). Is Justified True Belief Knowledge? Analysis 23 (6):121-123.

Moretti, Franco (2000). "The Slaughterhouse of Literature." Modern Language Quarterly, 61, no 1: 207-27.

Ross, W. D. (1930). The Right and the Good, Oxford: Oxford University Press, 1930.

Sloman S. A., Fernbach P. (2017). The Knowledge Illusion: Why we Never Think Alone. New York, NY: Riverhead Books.

Snow, C. P. (1959). The Two Cultures and the Scientific Revolution. New York: Cambridge University Press.

"The Artificial Intelligence and Data Act (AIDA) – Companion document." Canada.ca. Retrieved 2023-03-13 from https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document#s6.

Underwood, Ted (2019). Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press.