9

A Canon Is a Blunt Force Instrument: Data Science, Canons, and Generative Frictions

Aviva de Groot,[1] *George Fletcher,*[2] *Gijs van Maanen,*[3]
Akrati Saxena,[4] *Alexander Serebrenik,*[5] *and Linnet Taylor*[6]

Introduction

Spatially close, though worlds apart. The contributors to this commentary – 'we'; 'us' – conduct research and teach on data and technology-related issues at three Dutch universities. Some of us work at the same departments and teach in the same programmes. We bump into one another during our daily commutes and replenish our energy levels with the help of the same coffee machines after our lectures. We talk, and sometimes even discuss our research with one another. But do we also understand each other? What would that even mean? When we talk about 'data', do we talk about the same thing? Is that even necessary? What does 'science' for each of us entail? What does this mean for the education we collectively provide? What is the direction – scientifically, ethically, politically, professionally, commercially – the Bachelor programmes we are all involved in head toward?

National science policy in the Netherlands, as well as at the level of universities themselves, tends to prioritize in various ways computer and computational sciences over the social sciences and humanities (Taylor et al, 2023). We feel that the oppositions that are produced and reinforced through such policies are both false and unproductive, and this collective uneasiness motivated some of us to initiate a conversation about what it would mean to think and work together. How do our academic lives 'hang together' (Mol, 2014) beyond our encounters near coffee machines in the hallways, and our names on the timetables the students would find when logging in to their university pages?

When asking these and many other questions, we realized that we lacked the language, a common vocabulary, not only to answer the questions with which we started, but also to ask them.

Not only did many of the key concepts used in our research and education – data, algorithm, ethics, ontology, law – mean and do different things for all of us, but concepts indispensable to some – for example, justice – would be non-existent in the disciplinary universe of others.

We therefore needed to take a step back and reflect on how to have a conversation without sharing a common language. Our provisional solution was to take what we dubbed as 'canonical objects' as the focal points in our discussions. We borrow the notion of the canon from literary criticism, where it is used to mean a body of literature that over time comes to be taught as defining a particular culture (Bloom, 1994). For this reason, the canon has also been the focus of decolonial critics, who argue that we should critically interrogate the hegemonic discourses of Western culture (Spivak, 1990).

Based on this notion, we started to analyse concepts which each of us consider conceptually stable enough in our different disciplines that they might be taught on a Bachelor's-level course. In other words, we took our disciplinary backgrounds and educational responsibilities as conversational starting points. Our roughly defined meta-question was how our disciplinary backgrounds produced different conceptions of the same terms, how these differences could be generative or problematic, and how our disciplines become invested in a particular interpretation?

What we called canonical objects is also strongly related to how some of us used and understood the notion of boundary objects. A classic definition of boundary objects is that these 'have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation' (Star and Griesemer, 1989). Boundary objects thus allow different 'social worlds' to work together without requiring them to be able to (completely) understand one another. *If* our canonical objects would indeed function like boundary objects, we would have to find out and explicate in what way we would be working together, and how these concepts help us do that.

As part of our exploration, we also include answers from the generative large language model (LLM) ChatGPT3.5. This LLM draws on Internet content, and therefore offers a generalized and social version of the canon, replicating the most common tropes about our chosen objects of study available online. We wanted to include these tropes as a demonstration of how interdisciplinarity will often, if not theoretically informed, mash up disciplinary perspectives in ways that produce artificial neutrality and consensus on issues of real incompatibility, friction, and tension. As such, we aim to sensitize the reader to the incentives present in our different academic environments to focus on the lowest common denominator in debates on socio-technical phenomena, and to minimize disagreement.

Furthermore, interesting both conceptually and practically, was and still is, our attempt to create some level of mutual understanding (Gadamer, 2014), potentially with the help of boundary objects whose functioning depends on a *lack* of mutual understanding. How does our attempt to foster understanding about how we hang together or not, change our collaborations? What does this attempt do to the canonical objects that we used as conversational lubricants? How, to put that differently, does discussion and explicating our disciplinary divisions, change our capacities to, for example, teach together? And subsequently, what are generative but also less and non-generative ways of *dis*agreeing with one another?

In this contribution we present the results of the conversation we have had so far about two canonical concepts: artificial intelligence ('AI') and 'trust'. Together we made a list of *potential* canonical concepts (see the Appendix) – so concepts that would be taught in a BSc/BA programme/course – and from this list picked two of those with the most multifaceted disciplinary usage to discuss here. Each of us was asked to briefly explain how from their (disciplinary) point of view the concept was understood and taught in our undergraduate programmes. These brief reflections are accompanied by statements about our own positionality (Harding, 1991; Haraway, 1991) in which each of us situates her/himself in the academic tradition in which they were educated. We have included these because we presumed that academic disciplines (and what have been termed signature pedagogies (Poole, 2009)) were and still are the key factors that influence the types of academic social worlds most of us live in. In the discussion we present some of the themes that emerged in our conversation, and that help to understand how our academic activities hang together – or not.

Studying and teaching AI

AI, data systems, and George[7]

> *You wouldn't believe how difficult it is to get rid of philosophy. Not to philosophise is still to philosophise, but without method, with a sort of naive brutality.*
>
> Gabriel Séailles 1904

'Intelligence' has historically played a critical role in the evolution of the study of data systems in the computing science (CS) and data science (DS) disciplines. It can even be argued that the rise of the concept itself tracks the birth of these disciplines; it is one of the core ideas which enabled the emergence of these disciplines in the 20th century. I briefly highlight two canonical roles here; canonical in the sense that they have informed what is traditionally accepted as part of the common culture of these disciplines and, in particular, that of the data systems research community.

The first is the orientation towards and interest in intelligence itself and in the particular ways in which the concept has been understood. Our conceiving of Intelligence, and preoccupation with realizing this understanding, is our intellectual inheritance of Leibniz (for example, his idea of mechanical rationality, as famously articulated as 'Calculemus!'), Boole (for example, the logical and probabilistic framings of his Universal Laws of Thought), and their many fellow workers in the laying of the foundations and scoping of CS and DS prior to the 20th century (C. Babbage, C. S. Peirce, R. Llull, and others). The logical, rational, mathematical, and mechanical framings of what counts as Intelligence articulated by Leibniz, Boole, and company (namely, that there is something called Intelligence, that it is a feature of brains, and that brains are computers) set the stage (and the boundaries) for the CS/DS discipline's understanding of what meaningful questions we might ask about the world. These framings permeate the vocabularies and worldviews we inhabit in our research literature and in the classroom. Boole's division between logical and probabilistic reasoning set the stage for the two traditional camps of the discipline of AI, the so-called 'neats', developing symbolic reasoning, and the 'scruffies', developing inductive (probabilistic, statistical) reasoning. This is a structuring which still informs much of the work of AI and the related study of machine learning (ML).

The second is how this conceiving of and preoccupation with Intelligence is normalized (and thereby forgotten) in the broader field of computing and data science. The history of AI is very much the history of computing and data science, and within this the study of data systems; techniques and viewpoints developed first in AI eventually make their way into the rest of CS and DS as techniques and viewpoints in the day-to-day toolkits of researchers in areas beyond AI. My own area of study is data systems, the systems for the care, protection, and effective use of collections of data. A central aspect of any data system is the languages for interacting with the data collection, for articulating information needs. At the heart of these languages are capabilities for expressing the logical structure of data (for example, finding connections in a social network between people and the places they work) and the statistical structure of data (for example, determining the average salary at each such workplace). These capabilities are so mundane that we no longer note (or rather, are essentially completely unaware of) the intellectual roots of this orientation to datafy the world and then to understand it through these capabilities. Yet the analyses enabled by these capabilities, these 'universal' capabilities of 'thought' articulated by Boole and company, are at the base of the current developments in AI and ML.

Of course, canons evolve as fields evolve. There is building momentum within several areas of CS/DS towards a 'humanistic turn', placing people in the centre of our work, moving beyond our Leibnizian and Boolean notion of Intelligence and the harmful aspects of the shadow it casts over the study of data systems.[8] There are many signals of this turn. As an example, a recent experience I had in conversation with graduate students: students in the field are discovering and recommending to their advisers the work of early pioneers in the study of computing as a socio-technical phenomenon, such as Philip Agre (1995, 1997). Perhaps AI and data systems will gradually reconfigure their relationships in order to make more space in the canon for actual people (in place of the 19th-century Intelligence of AI), for creating space to study what it is we are doing together with data technologies and what we do to each other with these technologies and how we might do things differently and better, for people-centred notions of better.

AI, law, and Aviva[9]

Human intelligence is typically understood in law as the capacity to act in a 'rationally' informed way. It is of canonical legal concern in a very broad range of social situations, wherein what counts as rational depends on the specific legal interest in question. Law, for example, engages in what it sees as physiological approaches of cognitive agency for the assessment of human actors' accountability, culpability, or responsibility. For example, age, health (including mental) state, dis/ability are factored in rules around financial and other transactions, tort and criminal culpability, and the right to make medical decisions about oneself. The law's inherent responsiveness to social and political developments means it also gets lobbied to engage with, for example, novel neurological insights. Examples exist in the assumed moral and temporal 'shortsightedness' of the adolescent brain, and other neurological approaches of free will and autonomy that are debated in the field of 'neurolaw' (for example, Bigenwald and Chambon, 2019). Problematic historical examples of such responsiveness lead to the dark waters of craniometry, and to law's dealings with flawed intelligence-related findings from the field of genetics (Beckwith and Pierce, 2018).

Less physiological approaches of intelligence, for example, exist in the law's framing of the capabilities of 'average persons' with regard to their legal, medical, bureaucratic, and lately, technological literacy, translating into legal notions on how much, and what kinds, of information individuals 'reasonably' need in order to participate responsibly in medical, administrative, and other decision-making. Law's focus here is on the ability of persons to act in their self-interest. Intelligence, in such framings, tends to attach to privilege and to notions of merit and desert. First, the knowledge, experiences, and information needs of less privileged groups typically do not inform the 'average person' benchmarks (de Groot, 2023). Second, persons in need of (state) support are not 'lawyered' into information positions that help them to act in their self-interest – quite the opposite (for example, Bridges, 2017; Eubanks, 2018).

In short, the law is interested in framing intelligent human agency in order to govern situations in which it matters morally to do so. It works with theories, assumptions, and benchmarks for this. As signalled in preceding paragraphs, the 'North-Western' traditions that I teach in have inherited problematic colonialist and other ideological thought with regard to their modelling, and mainlining, of the individual, rational, reasoning, intelligent subject (Wright, 2001; D'Souza, 2018; Moyn, 2019; Katz, 2020). Corresponding rights (to vote, to perform political functions, enter into financial relations, decide about one's own body, not be discriminated against in education and career, not to be 'wrongly manipulated' in commercial advertising) have suffered as a result.

These are important themes when looking at how the law approaches what is called '*artificial* intelligence' and how it chooses theories, assumptions, and benchmarks to work with. For example, law tends to let itself be informed by the tech industry's own framings around complexity and explainability to inform, for example, transparency and explanation obligations that should help the intelligent, autonomous subject act in their self-interest when they are affected by AI. I like to alert students to how historically, framings around *medical* complexity were used by physicians many decades ago to argue against informed consent regulations, and as I will argue under the 'trust' lemma the same arguments are used today to mystify 'AI' (Katz, 1984; de Groot, 2023).

More generally, historically problematic ideological alignments between law and the AI fields have been researched from various disciplinary angles (for example, Benjamin, 2019; Katz, 2020; Lepore, 2020; Broussard, 2023), and some in technology law are acknowledging the need to engage with these insights.[10] Teaching (from) this critical angle is not standard practice in law schools in the Netherlands however, which are generally known to be conservative. This stands to impoverish legal education on digital technologies while courses on this are slowly being added to Bachelor curricula. The mentioned multidisciplinary histories of AI can usefully inform our 'budding canonical' teaching on 'AI' moving forward, and hopefully incentivize future scholars towards useful interdisciplinary engagements.

AI, political philosophy, and Gijs[11]

Just as doing philosophy is not a coherent and easily identifiable and definable activity, answering the question of how philosophers in general, or *political* philosophers, define AI, is a task designed to fail. I will nevertheless try to present some generalizations that hopefully capture the work that some philosophers, especially those working on politics, sometimes do. First, it is good to note that 'AI' and 'algorithms' are not always neatly distinguished from one another by philosophers. AI and algorithms have something to do with computers and machines that are engaged in activities that potentially relocate some capacities for action normally present in humans, to the realms of commerce and technology.[12] Whether or not this is problematic depends on how one thinks who or what is now in charge of our lives: the smart algorithms themselves, or the businesses who own them. Regardless of the specific direction an answer to this question takes, AI and algorithms pose or *could* pose fundamental problems to our human condition according to many philosophers (philosophers love problems).[13] Second, and third, what AI and algorithms are or do – apart from putting pressure on our human and societal capacity for action – is sometimes considered to be less important than the moral-political fact that they (could) present societal harms.[14] To put that differently: (political) philosophers often approach AI and algorithmically related problems through the moral and political lenses that they have been grinding the last 2,000 years, rather than starting their study of AI and algorithms in practice, and empirically. The conceptual and methodological toolkits philosophers use, offer them the ability to present fundamental moral-political critique of AI and algorithms (and sometimes also praise). They come, however, with the cost of presenting ambiguous and superficial analyses of the problems AI and algorithms pose in practice. One example to illustrate this point is the call to 'democratize' (the algorithms of) platforms such as Facebook because of the problematic character of Facebook's algorithm.[15] Arguments like these combine (a) philosophical theories on democracy, with (b) an idea that platforms use algorithms, and because of that, we should take back control, by (c) democratizing these algorithms. A possible limitation of such an approach is the conflation of extremely complex commercial ecosystems[16] and the presumed influence of 'algorithms' and 'AI' on what they do, potentially resulting in the wrong thing to take back control of (algorithms). It is sometimes a bit too easy and convenient to blame 'the mutant algorithm'.[17] Another difficulty concerns the indirect or implicit legitimization of platforms such as Facebook that goes hand in hand with the taking back control argument. To what extent is it of value to, indeed, take control over Facebook, and to what extent is democracy, or another normative theory, always applicable to supposedly problematic technology developments (see Dunn, 2005)?

AI, computer science curricula, and Alexander[18]

AI has a difficult relation with the traditional computer science curricula. For example, the model ACM Software Engineering curriculum (Ardis et al,

2015) does not mention it, while the 2020 Computing Curricula explicitly states that '[t]he study of artificial intelligence, an area of renewed interest, is not included in this report because an ACM/IEEE-CS sponsored curricular guideline does not currently exist'. Even curricula that do include courses on AI or related topics such as ML, tend to box them as a separate area of computer science endeavour not necessarily related to more traditional CS topics, such as algorithms, programming, or operating systems. AI courses within the traditional CS curriculum tend to avoid reflecting on the meaning of the terms 'artificial' or 'intelligence' in AI and use AI as an umbrella term covering several unrelated topics commonly associated with AI, such as knowledge representation, heuristic search algorithms, planning, ML, or neural networks. Is there a real difference between such computational tasks as sorting and planning, the first traditionally not being seen as an AI task, the second one seen as an AI task? Why is classification based on logistic regression not considered to be an AI task while classification carried out by a neural network is? In a way, traditional CS seems to treat AI as a metaphor: in the same way as Software Architecture describes main 'building blocks' of a software system, similarly to architecture describing actual building blocks, but is by no means subject to limitations of the physical world, AI is a study of tasks requiring 'intelligence', that is, tasks deemed to be too difficult for pre-AI approaches. However, metaphors have their limitations: revisiting the adage of Shimon Peres on pre-election polls, metaphors are 'like perfume – nice to smell, dangerous to swallow'. In other words, they have their charms, but one should be careful not to take them too literally. In a way, this triggers a fundamental question: does AI even exist as a coherent discipline?

Contemporary exposure of LLM-based techniques questions both the aforementioned vision as well as multiple elements of the traditional computer science curriculum. Do we still need to teach programming if LLMs can produce source code based on a series of natural language prompts? How would these models affect both the ways the software is going to look in the future and the ways it is going to be developed? The latter questions reflect the contemporary software engineering (SE) perspective on AI: similarly to philosophers, SE researchers are not interested in AI per se but rather in its impact on SE practices, trying either to understand how SE practices should be adjusted to address the challenges induced by the need to develop AI systems (SE4AI) or trying to understand how AI can be used or adjusted to support SE practices (AI4SE).

AI, computer science, and Akrati[19]

AI has a long and fascinating history in the field of CS. The origins of AI can be traced back to the 1950s, when computer scientists began exploring the idea of creating machines that could think like humans. Researchers such as John McCarthy, Marvin Minsky, and Claude Shannon laid the groundwork for AI with their groundbreaking work on topics such as logic and information theory. Around the 1970s, AI research focused on the development of expert systems, which were designed to mimic the knowledge and decision-making abilities of human experts in a particular field. Examples of these systems include MYCIN (Shortliffe, 2012), which diagnosed bacterial infections, and DENDRAL (Lindsay et al, 1993), which identified molecular structures. Later on, around the 1980s, the focus of AI research shifted towards ML, which involves training machines to learn from data.

When students study CS, they are offered a course on AI where they learn how to use programming, statistics, and robotics to design intelligent machines, also called expert systems or agents. CS students learn logic, programming, mathematics, data structure, algorithms, and software development in the starting years and then they learn AI through a combination of theoretical and practical coursework and also cover its applications in various fields. After acquiring theoretical knowledge, students start working on practical projects that involve developing AI systems. This involves using tools and libraries such as TensorFlow, PyTorch, and scikit-learn to create and train ML models. In these courses, students also learn how to handle large-scale datasets, how to process and clean data so that they can be used by AI systems. AI is a broad field, and students can specialize in various subfields of AI such as ML, Natural Language Processing (NLP), Robotics, and Computer Vision, and these areas are taught through specialized courses in Masters.

AI is a rapidly evolving field, and it is essential for students to keep up with the latest developments. They need to stay updated with new algorithms, tools, and libraries and continuously hone their skills through practice and experimentation. These days, universities are broadly promoting Bachelor and Master degrees in DS and AI, given the job market requirement. These courses focus on hands-on projects and practical exercise to give students a more practical understanding of the subject, as employers often look for candidates with practical experience and the ability to apply theoretical knowledge to real-world problems.

AI is not only used by the students who have a specialization or degree in it. These days, many people even from different backgrounds learn AI to apply it in their domain or learn it due to the big job market in this area. Most of these candidates focus on the practical aspect of AI. There are abundant online (most of them also free) resources and guidance that can help in gaining the knowledge and skills necessary to become successful AI practitioners. They mainly learn how to use AI APIs, data management tools such as SQL, Pandas, and if required then how to update the existing models based on the application requirements.

Back to our big question – 'What is AI?' In simple words, for computer and data scientists, AI is designing intelligent algorithms to automate predictive, repetitive, or decision-making tasks. As Alexander explained in the previous section, the concept of AI is closer to algorithms, as we design automation algorithms. AI has been used in many different areas from healthcare to finance to education and manufacturing. Most of the people designing AI systems are more involved in its practical application than its theoretical side.

AI, critical data studies, and Linnet[20]

Teaching data and AI governance to law students – one of the tasks I undertake as a teacher in critical data studies – means using a definition of AI that is usually very light on detail. The students are expected to work on legal problems rather than understand the technology, so the understanding of what AI-based systems are actually doing is extremely thin. The main effect of this is that even when students have been provided with greater technical detail, the legal disciplines tend not to encourage them to integrate that technical understanding with their analysis of legal and regulatory problems. Often, however, in order to make sense of a problem sufficiently to understand what to do to remedy it, it is necessary to have an idea of what features of the technology are generating that problem and how.

For a researcher in critical data studies, AI exists in layers. At one layer, it is (as Crawford and Joler have described, 2018), a 'system made of human labor', where material goods and infrastructures are integrated into what theorists have termed an 'assemblage' (Deleuze and Guattari, 1988) – a collection of heterogeneous elements working together to produce something with specific social effects – which then performs computational tasks. In the next layer, that of data, it consists of training data sourced from the social world, which theorists have defined as bodies (Floridi, 2016; Sandvik, 2023), social relations (Mhlambi, 2020; Tierra Común, 2021), or labour (Fuchs, 2014), concretized into factors of digital production. At another layer, that of statistical computation, those factors of production are transformed into moving parts of an analytical model, which weights different pieces of information differently and uses them in conjunction and opposition to each other to produce outputs which are then produced through an interface in a way that can make sense to particular groups of humans. Finally, there is the business layer, where those outputs are transformed into either knowledge or financial capital, and the assemblage earns its living.

For critical data studies, if we cannot understand who is involved in all these layers, we cannot understand the system and what it does. We also find the perspective of Stafford Beer useful (Beer, 2002), who, as a cybernetician, said that 'the purpose of a system is what it does', and not what it claims to do. What AI claims to do is readily visible from the literature produced by its developers and vendors. What it does, is more complex and requires situated research. If we compare research on AI governance to AI-based systems, we could say that most disciplines are using epistemologies that equate to symbolic reasoning – using hard-coded rules for what is relevant to the analysis, for choosing how to analyse that information, and for claiming that the resulting conclusions are valid. Critical data studies, like anthropology or some forms of philosophy, instead works in iterative interaction with the field: it proceeds through intuition and interaction, clustering inputs, seeking new patterns and correlations, and derives its validity from its underlying theory and methodological underpinnings, rather than by a stepwise explanation of process.

AI and ChatGPT 3.5

Prompt: Please explain what AI is, in 200 words, for an academic audience.

ChatGPT: AI refers to the ability of machines to simulate human intelligence and perform tasks that traditionally require human cognition, such as reasoning, problem-solving, perception, and learning. AI technologies encompass a broad range of techniques, including ML, NLP, computer vision, and robotics.

AI has a long history, dating back to the early 20th century, but recent advances in ML and big data have led to significant breakthroughs in AI research and application. AI is now used in a variety of fields, including healthcare, finance, transportation, and entertainment.

One of the most promising areas of AI is ML, which involves training algorithms to learn from data and improve their performance over time. This approach has led to significant advancements in speech recognition, image recognition, and NLP. Another important area is robotics, which involves the design and development of intelligent machines that can interact with the physical world.

While AI has great potential for improving human life and addressing complex societal problems, it also raises ethical and social concerns, including the potential for job displacement, bias, privacy infringement, and misuse. As such, the responsible development and deployment of AI technologies requires careful consideration of their social and ethical implications.

Studying and teaching trust

Trust, data systems, and George

The design of data systems is guided by several canonical concepts (guiding principles which underpin the field) towards realizing the dreams of Leibniz, Boole, Babbage, and company sketched in my earlier remarks on Intelligence. I will use one of these principles, the so-called ACID properties of transactional data systems, to illustrate a primary understanding of the concept of Trust in typical mainstream data engineering work.

The ACID properties govern the norms, the practical expectations, around data systems which manage transactions, that is, the updating and maintenance of a data collection as it evolves over time. ACID is typically taught in introductory courses on data management systems in CS and DS degree programmes. A classic example: when you make a purchase with your bank card, your bank balance must be correctly updated in the bank's database records for your bank account. ACID is an acronym denoting the four core principles of managing such transactions, to maintain the data system as the single point of truth: atomicity, consistency, isolation, and durability. Regarding Trust, 'durability' here means that an update should persist durably on some stable storage mechanism (for example, a hard drive) at the end of a transaction.[21] This means that even in the case of disruptions such as a system failure, the transaction will not be undone. In essence, a database system should be trusted to never forget data (in the absence of complete loss of the storage mechanism and all of its backup copies, such as in the case of a fire destroying the hard drive).

In the broader social contexts in which data systems are deployed, perfect remembering is often pathological behaviour. Humans forget, and for good reasons: personal growth, forgiveness, mental health. Societies forgive and forget, as reflected in laws governing the expungement of criminal records of minors and laws governing the right to erasure in digital systems. As Viktor Mayer-Schönberger has highlighted in his work, data forgetting is fundamental in a wide variety of social contexts (Mayer-Schönberger, 2009). This tension between the norms of data systems and the norms of society demands significant further study, especially in the data management research community.

Trust, law, and Aviva

A trust-framed look at law could start with the argument that laws are put in place to enable persons and institutions to interact in relative safety. Assuming that the need to trust arises when such safety cannot be assumed, law has an interest in trust relations that are formed to this end. The idea is that law serves society with democratically negotiated, knowable, foreseeable, and enforceable rules that allow subjects to rely on each other's and on institutional respect for public values and public order. Where law sees a role for itself, it operates from a standpoint of distrust, one could say: it defines trust-*worthiness*, *good* faith, *legitimate* expectations in light of the existence of their opposites. When things turn out badly, subjects can seek legal remedies and turn to legal institutions whose job it is to deal with cases of abused trust. Law Bachelor students are, for example, asked to memorize judgments about how X was reasonable to trust Y to not/act in a certain way in context Z, and that it was therefore reasonable that Y was condemned for abusing their position.

Lawmaking is a political activity that resonates with the outcomes of normative negotiations on all these points. And since legal rules work with semi-open norms, decisions on who and what counts as trustworthy are also responsive to those who populate the legal institutions, and research institutions, that interpret and further develop law. For trust in law itself, and in law-based societal rule, this matters greatly. Persons and communities whose values and interests are underrecognized in law and politics are less safe in their societies, and as a result they are trained in a rich variety of *dis*-trust.[22]

The subject of 'AI' is relatable to the aforementioned in many ways so I will briefly zoom in on a particular theme, namely insightfulness. First, to be able to investigate, interpret, track, and criticize how law co-shapes the normative landscape, that is, to gauge law's trustworthiness, law itself needs to be a sufficiently insightful process, practice, and paradigm. Law's progress has always depended on the hard work and activism of those who reveal law's biases, and continues to depend on them in 'AI-infused times' (Williams et al, 2022). The use of 'inscrutable' and/or experimental digital technologies in law-based policy, decision-making, and the judiciary, stands to complicate this challenging work. Second, it matters what standards law sets with regard to what constitutes trustworthy 'AI' and what kind(s) of insightfulness of AI practices are needed to assess this.

As mentioned in the 'intelligence' lemma above, in my teaching on the relations of insightfulness and trust in 'AI' contexts I like to discuss the longevity of the technological complexity argument. The argument was used in medicine around (and against) the paradigm shift from 'doctor knows best' to informed consent, that is, from trusted to trustworthy practices. Physicians who were uncomfortable with the shift at the time argued how medical knowledge was 'too technically complex' to ever be usefully understood by patients, and that trust should inevitably be invested in their medical authority. As Katz famously argued, however, other reasons for their explanation averseness hid behind the argument. Among other things he cites strong hierarchical cultures resulting in blind loyalty to medical elders; a medical lack of understanding of bodies and disease (especially for certain groups), medical chauvinism, and a widespread refusal to recognize the interests of patients (Katz, 1984). The arguments were supported by several equally problematic premises: that doctors can understand their patients well enough without engaging with their experience and standpoints, that 'medical knowledge' can be identified in isolation from other forms of knowledge, and that knowledge and decision-making about patients can proceed responsibly without patients' meaningfully informed participation in it. All these arguments and premises are recognizable in 'tech complexity' discussions about AI today, and in the technological lobby to keep law out of technological expertise-making while simultaneously calling for law to set up systems for accountability.[23]

To come back to the point of how groups' values and interests can be served better than those of others by law, attention should be called to how legal protections from wrongs and harms such as racism, marginalization, and discrimination have tended to fail less privileged communities well before 'AI'-driven decision-making became a widespread practice. One factor in why this is so is that law typically demands of individuals to prove how they were wronged in comparison with their peers, which is not possible when the abuse is systemic. In light of how 'AI' has been shown to exacerbate harms of this kind on a large scale, a progressive legal move could be to reverse burdens of evidence in contexts where subjects have historical reason to distrust the persons, institutions, and technologies they (need to) interact with. Legal efforts towards 'trustworthy AI' that disregard law's failures of protecting people from what are by now canonical 'AI' harms, can be considered a canonical failure.

Trust, political philosophy, and Gijs

Trust is not a canonical concept I use in general in my academic work, nor specifically in my lectures.[24] If I, however, was pressured to answer the question of the role of trust in political philosophy/theory, my answer would be something along the following lines. One strand of research in political theory/philosophy that might be more inclined to incorporate trust into their research agenda is research into the relationships between platforms, media, and the public sphere. For researchers working on these topics, democracy is often understood in a deliberative and knowledge-oriented endeavour where citizens, on the basis of publicly available information, discuss matters of public concern, as a means to reach agreement on how to deal with them. How information is shared with citizens, how the media works, how democratic debates and discussions are organized, and how truth is being produced in these processes are, therefore, key research topics. As part of such research on media and democracy, citizen's trust in both the facts they receive via the media, the media themselves, and each other, are topics to conduct research on.[25] Questions of trust understood from this angle, have received a significant boost in attention after the Cambridge Analytica scandal that showed how easy it is for big tech companies, through the manipulation of platforms/media, to affect the political decision-making process.[26] Trust, in other words, can become a topic of concern for political philosophers interested in technology but is then often considered to be part of a bigger puzzle on the functioning and quality of our democratic 'public spheres'. Scholars working in different subfields, or scholars like me who are not sure what to think of the ideal of the public sphere in the first place,[27] do not necessarily engage with 'trust'.

Trust, computer science, and Alexander

In computing, the topic of trust is predominantly discussed in the context of computer security: for example, a trusted computing base is a collection of hardware and software components deemed to be crucial for security of the entire system (Department of Defense, 1985), while a 'root of trust' is a 'tamper-resistant element in a digital system that can always be trusted, and therefore can be depended on as the root of all trusted operations' (Rowland and Karch, 2022). This interpretation of trust suggests that trust is being seen as a property of a system; it is aligned with the notion of trustworthiness – components that have met criteria related to correctness, safety, quality of service, security, and privacy have been deemed trustworthy (Hasselbring and Reussner, 2006). Alternatively, trust is associated with a reliable party certifying trustworthiness of the software: for example, macOS High Sierra *Trust* Store contains *trusted* root certificates,[28] issued by certificate authorities trusted by Apple, and subsequently by IT administrators of macOS High Sierra.

Another line of thought in computer security sees trust as a relation between communicating components (Garlan et al, 2009; Gai et al, 2010; also see more recent discussion of this topic in the context of the blockchain technology in Bryant and Saiedian, 2022) rather than property of an individual component; components then are seen as representing humans and performing series of simple actions on their behalf such as sending and receiving messages or performing computations. An example of this line of thought is the zero-trust security popularized by Google (Ward and Beyer, 2014), that, for example, states that 'a device that has not been updated with a recent OS patch level might be relegated to a reduced level of trust'. While presence of multiple often incompatible definitions of trust within the security community has been recognized as problematic, the common denominator seems to be that trust is absolute, established through complex hardware or cryptographic protocols and conflated with them.

Human–Computer Interaction (HCI) and, more recently SE, consider trust as a relation necessarily involving humans: for example, HCI positions itself in the space between individuals and software they interact with, while SE tends to take perspective of management science and study trust between members of software development teams. Finally, security aspects and trust as an interpersonal relation come together in security-related studies of topics related to deception and fraud in computer-mediated communication, for example, phishing.

Propagation of the AI techniques calls for rethinking the notion of trust and trustworthiness, triggering attention of researchers, practitioners, and policy makers to the topic of trustworthy AI, requiring AI to be lawful, ethical, and robust 'from a technical perspective while taking into account its social environment'.[29] In particular, robustness requires rethinking the aforementioned correctness, safety, quality of service, security, and privacy. The notion of correctness is particularly interesting: how can one distinguish inherent imprecision of AI techniques from bugs introduced in their implementation? Moreover, as more data or better AI techniques become available, answers produced by those techniques might change, requiring redefinition of the notion of correctness, trustworthiness, and ultimately truth.

Trust, data science, and Akrati[30]

In data science, when we talk about trust, it mainly refers to whether you can trust the output of the algorithm or an automated system. It works both ways, for scientists/researchers, it means to design systems that they can trust, and at the same time can be trusted by the users. To design trustable systems, data scientists consider trust at several steps, including:

- *Trustable data*: For computer scientists, trust in data is a critical consideration. Sometimes the data are generated due to biases in the system and cannot be used to train an AI model. For example, job hiring AI systems (Lewis, 2018) might have bias for female candidates. Beside this, data might be incomplete or manipulated (either intentionally or unintentionally), and this can have significant consequences. Therefore, computer scientists need to ensure that the data they are working with are accurate, reliable, and trustworthy.
- *Trustable algorithms*: In AI, we design algorithms/models to make decisions or recommendations, and in these cases, it is important to ensure that the algorithms are fair, transparent, and unbiased (Kleinberg et al, 2018). This means testing the algorithms rigorously and verifying that they are producing results that are consistent with expectations.
- *Trust in security*: Data scientists use databases to store the data and it is important to ensure their security from attacks. If the data storage is not secured, then it might risk users' data and they will not trust this system. Security measures can include things such as encryption, authentication, and access controls.
- *Trust in technology*: Finally, computer scientists need to consider trust in the technology itself. This means ensuring that the technology is reliable, robust, and performs as expected. It also means considering the ethical implications of the technology and ensuring that it is being used in a responsible and ethical manner.

By ensuring trust in these steps, data scientists can help ensure that their systems are trustable, reliable, accurate, and beneficial to society. However, even if a trustable system is designed, the main question is how to convince users that the system is trustable. For example, MYCIN was developed in the 1970s to assist medical professionals in diagnosing bacterial infections and recommending appropriate antibiotics to people. The success rate of MYCIN for prescribing correct medication was 65 per cent, that was higher than physicians who had an average success rate of 55 per cent (Liang, 1988). Despite its impressive accuracy and potential usefulness, people were hesitant to fully trust MYCIN for several reasons, including lack of transparency, limited knowledge base, and psychological factors.

The main question for data scientists is why people do not trust AI and how we can make people trust it. For the former question, the reasons are that there are several examples where AI has failed to qualify to be trustable, such as the hiring system (Raghavan et al, 2020), health care (Parikh et al, 2019), criminal identification (Angwin et al, 2016; Snow, 2018), facial recognition (Inioluwa and Buolamwini, 2019), translation systems (Prates et al, 2020), and so on. For later questions, AI scientists should focus on designing explainable systems with accountability, transparency, and ethical considerations. Beside these, educating people about AI systems and involving stakeholders in the design process might help people trust AI systems.

Trust, critical data studies, and Linnet

For critical data studies, the use of the terminology of 'trust' is a reliable sign that something is getting interestingly out of control. This is particularly true when the use of the term is purposely kept 'vague on who is to trust and whom, or what, is to be trusted'.[31] Keyes (2021) argues that current conceptualizations of trust in technology governance assume it is simple, transactional, and can be created to fit the needs of technology firms, and conversely that mistrust is defined as a flaw in users which must be fixed through rational explanation.[32] As such, it seems to be a continuation of the 'deficit model' of science communication, where people need to be improved and modernized so that they become more rational and appreciative of what science is doing (Simis et al, 2016). Trust, in this framing, can be understood as a companion to legal certainty for businesses: something that must be engineered and standardized so that innovation can proceed.

Conversely, there are more complex forms of trust reflecting subjective experience that are also interesting for scholarship of technology and its governance. Trust, for instance, as a dynamic component of relations between people and powerful entities such as corporations and public institutions. Here, trust in technology is functionally a subset of trust in power: people will trust to the extent that their interests are represented, and to the extent that the institutions they are trusting are accountable to them. Just as there would be no love poetry if love were always requited, trust would not enjoy the place it does in technology governance discourses if technological power were generally used in unproblematic ways. Therefore, as with democratic politics, the value of trust is frequently in the ways in which it is withheld.

'Trust' is most useful to critical data studies as a rhetorical signal that someone is in a position to behave badly at scale. When we hear the language of trust from regulators, rather than philosophers, we should ask what we are being asked to look away from. This is because most of the architectures that exist to prevent bad governance – electoral politics, the regulation of markets and businesses, research integrity codes, and most of public law, are designed to prevent people from having to trust the powerful. Critical data studies begins from the assumption that it is unlikely that technology is going to be the one phenomenon where this logic does not apply.

Trust and ChatGPT

Prompt: Please explain what trust is, in 200 words, for an academic audience.

ChatGPT: Trust is a fundamental concept in social sciences and humanities, and it is generally defined as the belief that an individual, group, or institution will behave in a reliable and predictable manner. Trust is a complex phenomenon that involves cognitive, emotional, and behavioural components, and it can be studied at different levels of analysis, from individual relationships to entire societies.

In interpersonal relationships, trust is often built through repeated interactions that demonstrate consistency, honesty, and integrity. When individuals trust each other, they are more likely to engage in cooperative behaviours and take risks together. However, trust can also be easily broken by betrayals, deception, or incompetence, and its restoration often requires significant effort and time.

In organizational contexts, trust plays a crucial role in shaping the behaviour of employees, customers, and stakeholders. Organizations that are perceived as trustworthy are more likely to attract and retain talent, build long-term relationships with customers, and generate positive reputations. Trust is also important in the context of institutions, such as governments and legal systems, which rely on the trust of citizens to maintain legitimacy and social order.

Trust has been studied from different theoretical perspectives, including social exchange theory, institutional theory, and social identity theory, and its mechanisms and consequences continue to be a subject of active research and debate in social sciences and humanities.

Discussion

The roughly formulated meta-questions with which we started this conversation were concerned with the relationships between our disciplinary backgrounds, the canonical concepts we choose, and the potential frictions emerging from the interaction of these different understandings, meanings, and worlds. At least three different themes pop up that relate to these questions, and that are worth explicating here.

On disciplinary straitjackets, and educating (in) DS

Many years ago, already, in a seminar series conveniently titled 'Discipline and Place' that one of us attended,[33] someone described academic disciplines as 'comfortable straitjackets'. Our discussion shows and confirms that the straitjackets we are all stuck in, determine how we (a) treat and (b) understand the canonical terms we chose to discuss. For legal scholars, for instance, 'intelligence' is directly related to human agency for reasons that have to do with the inherently normative character of the discipline of law. For data scientists, the concept of 'trust' is often understood as a response to the output of algorithmic or automated systems. It can even be an input to computing systems, rather than an output, because it can be defined as a quality possessed by infrastructures which are then used for computing processes. Researchers in critical data studies, by contrast, see the concept as a sign that something is going horribly wrong.

The roles of boundary objects

In contrast to discussing a particular disciplinary *understanding* of a canonical concept (b), some of us limited themselves to a description of how a term is treated in their discipline (a), without an accompanying substantive understanding of it. For political philosophers, concepts like algorithms, technology, and AI are often used interchangeably and primarily for the purpose of discussing several new and important moral-political problems the bad usage of such technologies could result in. In more traditional computer science curricula, AI is treated similarly: rather than presenting a specific analysis of the concept, it is being used as an umbrella term that incorporates all sorts of topics that in some sense relate to 'AI'. On the one hand, these descriptions might amount to a different way of *defining* and *understanding* things from within the perspective of that discipline ('one describes rather than defines'). On the other hand, and we take this to be more plausible, can these *treatment*s of especially 'AI' be interpreted as the instantiation of a boundary object ('AI') that fulfils various disciplinary functions. It allows philosophers, for instance, to connect themselves to new problems relating to 'technology' and by doing so further their field. Or, which is also not unimaginable, it helps to secure funding for research into data and technology-related research. What is key here, is that these usages of AI do not necessarily have to be accompanied with a full-fledged definition, conception, or understanding of 'AI'. It is all pleasantly fuzzy.

Tensions between disciplines, boundary objects, and education

Next to productive interactions resulting from the construal of boundary objects situated in between the computational fields and disciplines and the more social scientific and humanities ones, we did also notice destructive tendencies related to our situatedness in disciplines, and the need to connect to others. Such tensions manifest across a wide range of academic activities related to the way we build our fields, for instance when we found journals, publish, organize conferences, teach, or form research groups and institutes. To look at a few of these in more detail: when we seek support for our work, we are usually expected to do so under the auspices of a particular discipline, which determines how funders will read our proposals. In what is often a check-box exercise for reviewers, interdisciplinarity (where it is a criterion) must be identifiable, measurable, and calculable. Conversely, so must disciplinary affiliation. In the EU's Horizon grant programmes involving technology, for instance, there is often tacit signalling as to which disciplines are expected to participate in a particular call, and any proposal involving an unexpected mix of disciplines or the placement of a given researcher in an unconventional role must be clearly flagged and justified, and will usually raise scepticism from reviewers.

Conferences focusing on boundary objects such as 'AI' are becoming to some degree interdisciplinary, perhaps most of all the ACM's FAccT series on fairness, accountability, and transparency in computing. These latter series of conferences, however, are characterized by problems of different understandings of core concepts such as power, ethics, and bias (Ganesh et al, 2020). Moreover, they surface different disciplinary understandings of what constitutes ethical funding practices, with the norm for computing sciences being close collaboration with industry due to their predominantly applied nature and fluid boundaries with commercial technology providers, and the norm for social science and humanities disciplines being to keep much more distance from it for fear of ideological capture (Young et al, 2022). For the computing sciences, keeping up with developments in, and the needs of, industry, is a mark of good research. For the social sciences and humanities, influence by industry is closely controlled through research ethics codes, and ties with industry actors are scrutinized (though paradoxically also increasingly encouraged due to funding constraints) by universities.

Teaching across disciplines also surfaces these frictions. Students in DS are often, though not always, on a path towards careers in industry, and courses are designed to prioritize applied knowledge. The requirement to follow a course in ethics while doing a degree in DS is frequently seen as introducing an alien way of thinking into an otherwise coherent degree programme, where students being trained to think in a positivist paradigm must suddenly learn, often for the first time, to engage with critical scholarship on technology. Empirically, teachers have found that the introduction of deontological modes of reasoning about ethics, in particular, is hard for many students from an exact science background, who tend to find a consequentialist (even utilitarian) perspective most intuitive (Taylor and Dencik, 2020).[34] This is perhaps one of the most significant, but almost entirely unscrutinized, fault-lines between computing sciences and the social sciences and humanities: the assumption of a consequentialist rationale for both research goals and the application of technology is such a basic underpinning for the computing sciences that it is difficult for students to imagine their way out of it – and for meaningful forms of ethical scrutiny to be applied to computing and data science projects (Metcalf and Crawford, 2016).

On (de)professionalizing boundary objects

Our contributions suggest that the incorporation of canonical concepts in disciplines, as a means to develop a discipline internally or to connect it to other fields, disciplines, or funding streams, should not be left uninterrogated.

For starters, as highlighted in different ways by several of us, the incorporation of generalized concepts without careful scrutiny – a standard practice across our fields – could result in a limited grasp of the problems researchers should deal with. Platforms, for instance, cannot be equated with algorithms, which is the reason critical data studies researchers deconstruct and expose the various layers that together make 'AI'.

Going slightly beyond such epistemologically related limitations, we also notice political economic worries with respect to the treatment of these and other canonical concepts in our disciplines. The pragmatic cross-disciplinary agreements on boundary objects we have explored earlier have the effect of making it possible to stabilize, and therefore also professionalize, the process of studying, teaching, and monetizing them. There has to be some coherence in our DS programme in how we use concepts, because, among other things, one has to find consensus on definitions when applying for one's interdisciplinary grant proposals. As such, we might see technological boundary objects in the world of data and AI *becoming* boundary objects through their links to industry and public and private funders. The tech-oriented boundary objects stabilize and reinforce relationships and collaborations between education and industry to such an extent that our capacities to come up with, and interject in our classes, objects and languages that are oriented towards other-than-industrial languages and worlds has been reduced.

Disciplinary affiliation, and the security offered by these straitjackets, has a price, and attempts to critically interrogate the fundamental assumptions of our disciplines puts us out of step with colleagues with whom we could otherwise be collaborating on papers, education, and grant proposals.

An important characteristic of the boundary objects encountered in our canons, in sum, is the incentivization of disciplinary obedience geared towards the production of efficient and frictionless collaboration. By taking our canons as the starting point of our conversation, we acquired a better understanding of the boundary objects that help to communicate without really understanding one another. The resulting lessening of the comfort granted by our straitjackets will, we hope, increase the room we have for manoeuvre within and in between our fields, potentially resulting in less efficient, productive, and frictionless possibilities not directed towards disciplinary renewal, but a better world.

Appendix: canonical terms

- Trust
- Algorithms
- Governance
- Ethics
- Explanation
- Law
- Infrastructure
- Accountability
- Data
- Fairness
- Justice
- Causality
- Truth
- Efficiency
- Significance
- Ontology
- AI
- Politics
- Power
- Ambiguity
- Accuracy
- Probability

Notes

[1] Tilburg University, the Netherlands, aviva.degroot@tilburguniversity.edu

[2] Eindhoven University of Technology, the Netherlands, g.h.l.fletcher@tue.nl

[3] Tilburg University, the Netherlands, g.vanmaanen@tilburguniversity.edu

[4] Leiden University, the Netherlands, a.saxena@liacs.leidenuniv.nl

[5] Eindhoven University of Technology, the Netherlands

[6] Tilburg University, the Netherlands, l.e.m.taylor@tilburguniversity.edu

[7] I write from the perspective of a data language researcher, trained and working for the past two decades in the domain of data management systems, a subfield of the computing and data sciences. I have been on the faculty at a Dutch technical university for the past 14 years; prior to this I was an assistant professor at a West-coast American comprehensive public university. Before completing my PhD on topics in data integration (at a comprehensive public university in the American Midwest), I double majored in mathematics and liberal studies in cognitive science, writing my Bachelor's thesis on topics in cognitive science viewed through a social philosophy lens (at a comprehensive public university in the American South). As a child I lived between Japan, with extended family centred on Yonaguni Island, and America, with extended family centred in the Appalachian mountains of western North Carolina. I teach in Bachelor's and Master's programmes in computing and data science.

[8] For a recent discussion of examples of these harmful aspects and their historical roots, see Whittaker (2023).

[9] I am a postdoc in AI & Human Rights, a position I landed after a PhD in which I dissected several canonical legal 'explanation obligations' in light of how such obligations are seen to be hard to meet in the context of today's automated decision-making methods. Informed by historical accounts of both AI and law, and by earlier work experience in the field of legal aid, I am increasingly sceptical of such 'novelty' leaning problematizations. Working with differently trained researchers helps me to identify new iterations of old problems, and to think through how solutions can take a more holistic approach. In doing so I am confronted with how the separation of types of work, types of workers, and the power relations that attach to these separations poses obstacles for more inclusive work towards (global) social justice. But working this way can also complicate single-discipline work collaboration. Working with more singularly oriented legal scholars, I tend to (feel the need to) unsettle disciplinary understandings of, for example, 'well-established' fundamental values. Whether that happens also depends on colleagues' more personal standpoints. For example, differences in legal focuses and standpoints also tend to express in, and as, political colour, which influences the type of students a university attracts and the scholarship that is produced by its researchers. Adding more personal bios to articles seems like a good idea in light of this, especially in collective endeavours with one person per discipline. Aviva, 'the' lawyer? No, Aviva, 'a' lawyer. I have professional backgrounds in cabinet making, (SF and horror) filmmaking, and legal aid – where protecting persons against an ill-wishing, tech-happy State was core business.

[10] https://edri.org/what-we-do/decolonising-digital-rights/

[11] I studied history and (political) philosophy, and focused on democratic theory, migration/multiculturalism, feminist/new materialist theories, and science and technology studies (STS). When looking for a PhD position I ended up at Tilburg Law School where I joined an administrative law project on 'citizen-friendly data communication', and where it was my job to reflect on ethical principles that could help to guide such communicative processes. Being uncomfortable with doing 'ethics' in that way, I decided to do an empirical analysis of open government and open data policies/practices in the Netherlands instead. Methodologically speaking, this amounted to a mixture of STS/ethnography and political theory. Afterward, during my postdoc at a private law institute elsewhere, I combined literature from political theory and (political) economics in a research project on collective data governance. I, in other words, consider myself to be a disciplinary migrant that travels to and tries to make sense of the academic debates and disciplines that help to make sense of a problem. I teach in the Bachelor's Data Science progamme in Tilburg/Eindhoven where

I try to make the students receptive to ethical-political problems related to 'technology', and draw from a mixture of political philosophy/theory, data ethics (broadly conceived), critical data studies, and STS. For this contribution, I primarily thought about how from a political philosophical point of view, the two canonical concepts would be approached, and drew from more accessible/introductory texts on the matter (of which there are few).

[12] Though, for example, Risse in the beginning of his book distinguishes between algorithms, machine learning algorithms, 'specialised AI', and 'general AI', the concepts are often used relatively interchangeably in the rest of the book (Risse, 2023).

[13] Coeckelbergh wrote one of the few accessible overviews of the relationship between political philosophy/theory and AI/algorithms, and prioritized the problems that different technologies such as recommender systems pose to, for instance, human freedom (Coeckelbergh, 2022).

[14] See both Risse (2023) and Coeckelbergh (2022).

[15] Drawing from critical/Marxist work on the 'commons', Thijs Lijster makes this argument (Lijster, 2022). Ugur Aytac also presented a similar argument at the Mancept workshops 2022: https://sites.manchester.ac.uk/mancept/mancept-workshops/programme-2022-panels/digitaldemocracy/

[16] See, for example, the work of Anne Helmond: www.annehelmond.nl/publications/

[17] As Boris Johnson labelled the Ofqual algorithm used to predict the grades of students that were unable to do their final exams due to the COVID pandemic: www.bbc.com/news/education-53923279

[18] I am a classically trained computer scientist that has slowly migrated to software engineering and, specifically, to studies of human and social aspects of software engineering. While my expertise is in software engineering, in the texts, I am trying to present a broader view of CS in general. This is, of course, a challenging endeavour due to the variety of topics studied in CS and different perspectives CS disciplines have on the topics discussed in this chapter.

[19] I am a computer science engineer by training and then extensively worked on network science and data science. I write from the perspective of a computer scientist and especially, a data scientist. Currently, I am working as an assistant professor at Leiden Institute of Advanced Computer Science, Leiden University, the Netherlands. I have taught data mining and related courses to Bachelor's and Master's students and written about how students learn AI and then what AI means to them.

[20] I am a human geographer and critical data studies researcher working in the law faculty at Tilburg University. I have taught in International Development Studies, Data Science, and international governance at Bachelor's and Master's levels.

[21] The others stand for, respectively, atomicity, consistency, and isolation.

[22] The extent to which lawmakers and legal institutions care about this depends. When a critical mass of legal subjects lose trust in rule-based society, law, and the powers that law is assumed to keep in check, both stand to lose legitimacy – and power.

[23] See, for example, the call to establish licensing systems for building 'cutting edge models' and establishing liability for foreseeable and preventable harms: https://managing-ai-risks.com/

[24] It is not extensively being discussed by either Risse or Coeckelbergh, and also the 'tech ethics' handbooks do not include it in their indexes (van der Poel and Royakkers, 2011; Taebi, 2021; Nyholm, 2023). An exception is a chapter on Kantian ethics in a recently published technology ethics anthology (Myskja, 2023).

[25] For example, Farrell and Schwartzberg, 2021.

[26] Coeckelbergh does discuss such problems in chapter 4 of his book ('Democracy: echo chambers and machine totalitarianism'), though does not give a substantive analysis of the role of trust in these matters.

[27] See, for example, the collection of reflections on Habermas' concept in Calhoun (1992).

[28] https://support.apple.com/en-us/HT208127

[29] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[30] I am writing about what trust means for data scientists and for people using these AI systems.

[31] https://algorithmwatch.org/en/trustworthy-ai-is-not-an-appropriate-framework/

[32] https://reallifemag.com/standard-evasions/

[33] Gijs participated in the mentioned seminar. A course overview can be found here: https://studiegids.universiteitleiden.nl/en/courses/101290/discipline-and-place-in-the-social-sciences-and-the-humanities

[34] Consequentialism and utilitarianism are closely related strands in the philosophical literature – and especially, moral and political philosophy – that evaluate the (moral) rightness of our actions on the basis of the consequences they would have. See https://plato.stanford.edu/entries/consequentialism/. This can be contrasted with theories that prescribe that actions should be in accordance with particular moral rules or principles ('You should never lie, regardless of the consequences of your (in)action').

References

Ardis, M., Budgen, D., Hislop, G. W., Offutt, J., Sebern, M., and Visser, W. (2015) SE 2014: Curriculum guidelines for undergraduate degree programs in software engineering. *Computer*, 48(11): 106–109.

Agre, Philip. (1995) The soul gained and lost: artificial intelligence as a philosophical project. *Stanford Humanities Review*, 4(2): 1–19.

Agre, Philip. (1997) 'Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI'. In: Geoffrey C. Bowker, Susan Leigh Star, William Turner, and Les Gasser (eds) *Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide*, 131–158. Hillsdale, NJ: Erlbaum.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016) 'Machine bias'. ProPublica, 23 May. Available at: www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Beckwith, J. and Pierce, R. (2018) 'Genes and Human Behavior: Ethical Implications'. In: Robert T. Gerlai (ed) *Molecular-Genetic and Statistical Techniques for Behavioral and Neural Research*, 599–622. San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-804078-2.00025-8.

Beer, S. (2002) 'What is cybernetics?' *Kybernetes*, 31(2): 209219.

Benjamin, R. (2019) *Race after Technology*. New York: Polity Press.

Bigenwald, A. and Chambon, V. (2019) 'Criminal responsibility and neuroscience: no revolution yet'. *Frontiers in Psychology*, 10: Article 1406. https://doi.org/10.3389/fpsyg.2019.01406

Bloom, H. (1994) *The Western Canon: The Books and School of the Ages*. New York: Harcourt Brace & Company.

Bridges, Khiara M. (2017) *The Poverty of Privacy Rights*. Stanford, CA: Stanford University Press.

Broussard, M. (2023) *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. Cambridge, MA: MIT Press.

Bryant, B. D. and Saiedian, H. (2022) Key challenges in security of IoT devices and securing them with the blockchain technology. *Security and Privacy*, 5(5).

Calhoun, C. (ed) (1992) *Habermas and the Public Sphere*. Cambridge, MA: MIT Press.

Coeckelbergh, M. (2022) *The Political Philosophy of AI: An Introduction*. Cambridge: Polity Press.

Crawford, K. and Joler, V. (2018) 'Anatomy of an AI system'. Available at: www.anatomyof.ai

D'Souza, R. (2018) *What's Wrong with Rights? Social Movements, Law and Liberal Imaginations*. London: Pluto Press.

de Groot, A. (2023) Care to explain? A critical epistemic in/justice based analysis of legal explanation obligations and ideals for 'AI'-infused times. PhD thesis. Tilburg University, Tilburg.

Deleuze, G. and Guattari, F. (1987) *A Thousand Plateaus: Capitalism and Schizophrenia*. Minneapolis: University of Minnesota Press.

Department of Defense (1985) 'Department of Defense trusted computer system evaluation criteria', DoD 5200.28-STD

Dunn, J. (2005) *Setting the People Free: The Story of Democracy*. London: Atlantic Books.

Eubanks, V. (2018) *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Farrell, H. and Schwartzberg, M. (2021) 'The Democratic Consequences of the New Public Sphere'. In: Lucy Bernholz, Héléne Landemore, and Rob Reich (eds) *Digital Technology and Democratic Theory*, 191–218. Chicago, IL: University of Chicago Press. Chicago Scholarship Online. https://doi.org/10.7208/chicago/9780226748603.003.0008

Floridi, L. (2016) On human dignity as a foundation for the right to privacy. *Philosophy & Technology*, 29(4): 307–312. https://doi.org/10.1007/s13347-016-0220-8

Fuchs, C. (2014) Digital prosumption labour on social media in the context of the capitalist regime of time. *Time & Society*, 23(1): 97–123. https://doi.org/10.1177/0961463X13502117

Gadamer, H.-G. (2014) *Truth and Method*. London: Bloomsbury Academic.

Gai, X., Li, Y., Chen, Y., and Shen, C. (2010) 'Formal definitions for trust in trusted computing'. UIC/ATC Workshops 2010: 305–310.

Ganesh, M. I., Dechesne, F., and Waseem, Z. (2020) Two computer scientists and a cultural scientist get hit by a driver-less car: a method for situating knowledge in the cross-disciplinary study of FAT in machine learning: translation tutorial. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp 707–707.

Garlan, D., Allen, R., and Ockerbloom, J. (2009) Architectural mismatch: why reuse is still so hard. *IEEE Software*, 26(4): 66–69.

Haraway, D. (1991) *Simians, Cyborgs and Women: The Reinvention of Nature*. New York: Routledge.

Harding, S. (1991) *Whose Science? Whose Knowledge? Thinking from Women's Lives*. New York: Cornell University Press.

Hasselbring, W. and Reussner, R. H. (2006) Toward trustworthy software systems. *Computer*, 39(4): 91–92.

Inioluwa, R. D. and Buolamwini, J. (2019) Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. New York: Association for Computing Machinery, pp 429–435. https://doi.org/ 10.1145/3306618.3314244

Katz, J. (1984) *The Silent World of Doctor and Patient*. Baltimore, MD: Johns Hopkins University Press edition, 2002.

Katz, Y. (2020) *Artificial Whiteness: Politics and Ideology in Artificial Intelligence*. New York: Columbia University Press.

Keyes, O. (2021) 'Standard evasions'. *Real Life*, 30 August. Available at: https://reallifemag.com/standard-evasions/

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018) Algorithmic fairness. *AEA Papers and Proceedings*, 108: 22–27.

Lepore, J. (2020) *If Then: How the Simulmatics Corporation Invented the Future*. New York: Liveright Publishing.

Lewis, N. (2018) 'Will AI remove hiring bias?' *Strategic HR Review*, 11 November. Available at: www.shrm.org/topics-tools/news/talent-acquisition/will-ai-remove-hiring-bias

Liang, T. P. (1988) Expert systems as decision aids: issues and strategies. *Journal of Information Systems*, 2(2): 41–50.

Lijster, T. (2022) *Wat We Gemeen Hebben: Een Filosofie van de Meenten*. Amsterdam: De Bezige Bij.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., and Lederberg, J. (1993) DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, 61(2): 209–261.

Mayer-Schönberger, V. (2009) *Delete: The Virtue of Forgetting in the Digital Age*. Princeton, NJ: Princeton University Press.

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

Mhlambi, S. (2020) From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance. *Carr Center for Human Rights Policy Discussion Paper Series*, 9: 31.

Mol, A. (2014) *The Body Multiple: Ontology in Medical Practice*. Durham, NC; London: Duke University Press.

Moyn, S. (2019) *Not Enough: Human Rights in an Unequal World*. Cambridge, MA: Harvard University Press.

Myskja, Bjørn K. (2023) 'Technology and Trust – A Kantian Approach'. In: Gregory J. Robson and Jonathan Y. Tsou (eds) *Technology Ethics: A Philosophical Introduction and Readings*, 122–129. New York: Routledge.

Nyholm, S. (2023) *This is Technology Ethics: An Introduction*. Hoboken, NJ: Wiley.

Parikh, R. B., Teeple, S., and Navathe, A. S. (2019) Addressing bias in artificial intelligence in health care. *Jama*, 322(24): 2377–2378.

Poole, G. (2009) 'Academic disciplines: Homes or Barricades?' In: Carolin Kreber (ed) *The University and its Disciplines: Teaching and Learning Within and Beyond Disciplinary Boundaries*, 55–57. New York: Routledge.

Prates, M. O., Avelar, P. H., and Lamb, L. C. (2020) Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32: 6363–6381.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020) Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp 469–481.

Risse, M. (2023) *Political Theory of the Digital Age: Where Artificial Intelligence Might Take Us*. Cambridge: Cambridge University Press.

Rowland, M. and Karch, B. (2022) 'A Review of Technologies that can Provide a 'Root of Trust' for Operational Technologies'. Sandia Report SAND2022–3755.

Sandvik, K. B. (2023) *Humanitarian Extractivism: The Digital Transformation of Aid*. Manchester University Press.

Séailles, G. (1904) 'L'enseignement secondaire et la philosophie'. In: *Éducation ou revolution*. Paris: Armand Colin.

Shortliffe, E. (ed) (2012) *Computer-based Medical Consultations: MYCIN* (Vol. 2). New York: Elsevier.

Simis, M. J., Madden, H., Cacciatore, M. A., and Yeo, S. K. (2016) The lure of rationality: why does the deficit model persist in science communication? *Public Understanding of Science*, 25(4): 400–414.

Snow, J. (2018) 'Amazon's face recognition falsely matched 28 members of Congress with mugshots'. American Civil Liberties Union. Available at: www.aclu.org/blog/privacytechnology/surveillance-technologies/amazonsface-recognition-falsely-matched-28

Spivak, G. C. (1990) The making of Americans, the teaching of English, and the future of culture studies. *New Literary History*, 21(4): 781–798.

Star, S. L. and Griesemer, J. R. (1989) Institutional ecology, 'translations' and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, 19(3): 387–420. https://doi.org/10.1177/030631289019003001

Taebi, B. (2021) 'Ethics and Engineering: An Ethics-Up-Front Approach'. In *Ethics and Engineering: An Introduction* (Cambridge Applied Ethics), 1–20. Cambridge: Cambridge University Press.

Taylor, L. and Dencik, L. (2020) Constructing commercial data ethics. *Technology and Regulation*, 1–10.

Taylor, L., van Maanen, G., de Ridder, J., Taebi, B., and Bruijnincx, P. (2023) Space to think. An analysis of structural threats to academic freedom and integrity. Amsterdam, KNAW. Available at: www.dejongeakademie.nl/publicaties/2495595.aspx

Tierra Común (2021) *Interventions for data decolonization*. Available at: www.tierracomun.net/en/home

van der Poel, I. and Royakkers, L. (2011) *Ethics, Technology, and Engineering: An Introduction*. Hoboken, NJ: John Wiley.

Ward, R. and Beyer, B. (2014) 'BeyondCorp: a new approach to enterprise security'. login Usenix Mag., 39(6).

Whittaker, M. (2023) Origin stories: plantations, computers, and industrial control'. *Logic(s)*, Issue 19.

Williams, P., Kuntsman, A., Nwankwo, E., and Campbell, D. (2002) 'Surfacing systematic (in)justices: a community view'. Available at: https://systemicjustice.ngo/wp-content/uploads/2023/06/SystemicJustice_Report_FINAL.pdf

Wright, S. (2001) *International Human Rights, Decolonisation and Globalisation Becoming Human*. London: Routledge.

Young, M., Katell, M., and Krafft, P. M. (2022) 'Confronting power and corporate capture at the FAccT conference'. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp 1375–1386.