A Framework for a Data Interest Analysis of Artificial Intelligence

Author: Gry Hasselbalch, working paper, in review, FirstMonday, 2020.
Contact: mediamocracy@protonmail.com

Abstract

This article makes a case for a data interest analysis of artificial intelligence (AI) that explores how different interests in data are empowered or disempowered by design. The article uses the EU High-Level Expert Group on AI's Ethics Guidelines for Trustworthy AI as an applied ethics approach to data interests with a human-centric ethical governance framework and accordingly suggests questions that will help rebalance conflicts between the data interest of the human being and other interests in AI design.

Keywords: artificial intelligence, big data, STS, VSD, data ethics, data interests, power, ethics by design, Trustworthy AI, human-centric, privacy, fundamental rights.

Introduction

Data flows. Data transforms. Data changes hands, bodies and containers. The next frontier in the age of big data flows is artificial intelligence (AI) systems that are developed to contain and make sense of large amounts of data and to act on that knowledge. As AI data systems become integrated into society, they also turn into centers of negotiations between different interests in data that join, appear and disperse in the data design of AI as a symptom of and condition for the distribution of agency and powers in society. This article argues for an exploration of data interests in AI to help make ethical choices in the development of AI.

In the late 2010s, the term "artificial intelligence" gained traction in public discourse. Business and technology companies started rebranding their big data efforts as "AI" (Elish & Boyd, 2018), and in policymaking AI became an item of strategic importance worldwide. In public and private sectors, decision-making processes were progressively informed by and even replaced by big data AI systems. Recommendation and personalization systems were profiling and analyzing our personal data and deciding for people what they see and read and with whom they engage online. Judicial risk assessment systems were looking for patterns in backgrounds of defendants to inform judges about who would be most likely to commit a crime in the future. Triage systems processed the medical and demographic history of patients to decide who would get a kidney. Moral decisions and choices were increasingly intertwined with AI systems' complex data processing, and accordingly interests in the data of AI as a resource to protect, share, acquire and to be empowered or disempowered by came together in efforts to direct the development of AI.

Against this background, an institutionally framed European AI agenda took shape with an emphasis on "ethical technologies" and "Trustworthy AI". This applied AI ethics approach was described in core documents and statements in a process that involved, among others, an independent multi-stakeholder expert group on AI ("The EU High-Level Expert Group on AI", AI HLEG) that in 2019 published the "Ethics Guidelines for Trustworthy AI".

Several other ethics guidelines and sets of principles for AI, such as these, were at the time created by various stakeholder interest groups. Most of themshown to share common themes (Fjeld et al., 2019; Floridi et al., 2018; Jobin et al., 2019; Winfield & Jirotka, 2018). Yet, while thematic convergence is indeed relevant when seeking out general global consensus on applied ethics approaches to AI, I want to argue that each guideline also has a unique point of reference and, importantly, is positioned within distinct cultural, social and legal frameworks.

The AI HLEG guidelines were framed with a human-centric approach to the development of AI: "The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed, deployed, used and monitored…" (p. 37). In the late 2010s this "human-centric" approach was a term and a theme that had emerged in policy discourse on AI with no common conceptualization other than an emphasis on the special role and status of humans (e.g. OECD, 2019, European Parliament, 2019). However, the AI HLEG's work with the ethics guidelines was from the outset framed in terms of a distinctive European agenda created to ensure the development of a European AI eco system (Hasselbalch, forthcoming). The guidelines were therefore also grounded in a European fundamental rights legal framework and the European data protection legal reform with an emphasis on the autonomy and dignity of the individual human being.

The policy idea that the design development of a socio-technical may informed by overarching ethical reflection, such as the human-centric one, calls attention to the moral qualities of technologies, that is, respectively their embedded "values" or their "politics". In the following, I describe AI systems as agents to which humans delegate the enforcement of the agency of different interests in society. Recognizing this type of delegated agency of interests allows us to develop and design with a human-centric ethical reflection and to make choices accordingly. I begin with an examination of the concept of "the human interest" and "data interests". The first premise for a data interest analysis is that the very design of a data system represents power dynamics that a technology may create by design "wonderful breakthroughs by some social interests and crushing setbacks by others" (Winner, 1980, p. 125). While an applied ethics approach, such as Value Sensitive Design (VSD), allows us to isolate the very design phase of a data system to analyze the way in which stakeholder values are negotiated by design, a Science of Technology (STS) framework treats the power (moral or political qualities) of technologies as dynamic concepts in constant negotiation with societal factors. This is the second premise of the article that considers the development of AI data systems as a component of society at large, evolving laws and policies, economy and culture. Here, I move on to the depiction of data interests in the AI HLEG's ethics guidelines. As the AI HLEG does not explicitly address data interests as such, I use a metaphorical analysis as a tool to illustrate the guidelines' position on how to resolve conflicts between different data interests. I suggest five themes in regard to data interests (data as resource, data as power, data as regulator, data as vision and data as risk) that should be considered in the ethical governance of AI.

This article is not concerned with the philosophical concept of AI, nor with the scientific aspiration to create human-level machine intelligence. It addresses the practical adoption of AI in society as data intensive systems and uses a specific applied ethics approach to the development of these systems as suggested by the AI HLEG. Therefore, the focal point is, above all, on machine learning, which is essentially a data processing system with different degrees of autonomy. By the late 2010s, amplified computer power and the vast amount of data generated in society had empowered machine-learning technologies to evolve and learn to recognize faces in pictures, to recognize speech from audio, to drive a car autonomously and to understand individuals when, for example, microtargeting services and information, etc. These were all practically applied, more or less autonomous systems of data processing adopted by companies and states to not only solve simple problems, analyze and streamline disparate data sets but to act in real time, sensing an immediate environment and to support critical human decision-making processes. As such, their ethical implications in regard to human agency and involvement arose from systems of distributed moral decision-making between humans and "non-human" agents.

The Human Interest in AI development

The human-centric approach to AI emerged as a shared emphasis in the policy debate on AI of the late 2010s and is generally associated with the well-being of the individual human being and the societies and environments of humans in the design, development and implementation of AI of (e.g. in AI HLEG B, 2019, p.9). As follows, a key emphasis is placed on AI design that prioritizes human involvement, human agency and human decision-making. In the European policy context, this is first and foremost associated with the autonomy, dignity and fundamental rights of the individual human being. As core concepts of the European Fundamental Rights framework that form part of the core principles of European law, the human-centric approach therefore here also goes beyond a concern with the design of AI systems only, compelling a socio-technical environment in which fundamental rights are implementable.

AI systems have been deployed in different societal sectors since the 1980s. They evolved from rule-based expert systems encoded with the knowledge of human experts that were applied in primarily human and physical environments into machine-learning systems, evolving and learning from big data in digital environments with increasingly autonomous decision-making agency and capabilities (Alpaydin, 2016). Despite the fact that machine learning to a certain extend cut out the human expert, it did not entirely exclude human involvement in the very data design of the system. On the contrary, the degree of an AI system's autonomy is, for example, prescribed by the human involvement in data processing from the definition of the problem, the collection of data and data cleaning to the training of the machine-learning algorithm (Lehr & Ohm, 2017). Lehr and Ohm refer to this human involvement in machine-learning processes as "playing with data". Machine-learning algorithms, as they state, are not "magical" black boxes with mysterious inner workings. In fact, they are the "complicated outputs of intense human labor—labor from data, scientists, statisticians, analysts, and computer programmers" (Lehr & Ohm, 2017, p. 717).

In this way we may associate the human interest in the data of AI in very practical terms as an involvement of human actors in the very data design, use and implementation of AI. In the AI HLEGs ethics guidelines the human-centric approach is for example spelled out in a particular attention to the interests of the individual human being and "human-in-command" and "human agency and oversight" components in the design and conditions for the development of AI. Though the preoccupation with the human interest in AI data design may also be considered more generally in terms of a critical reflection on the very status of human agency and AI agency. Responding to a conflict in the original aspirations of AI research to create respectively machines that thinks and understands by themselves or machines that "just" process information and solve problems for humans, John R. Searle famously argued in 1980 that strong AI has "little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking." (Searle, 1980, p.417). Today public discourse on AI carries traces of the first aspiration to build artificially intelligent agents comparable to the human agent and accordingly conceptions of the imminent potentials or threats of AI. For example, concerns regarding AI agents that replace the human labour force or the artistry and creativity of new AI systems represent an imagining of an autonomous out of human control artificial agent. As such, it may be argued as, Ellish & Boyd does, that this type of "magic" surrounding Artificial Intelligence also disempowers us in what we think we can do with AI (Ellish & Boyd, 2017). Along these lines, another "human-centric" call for action has been made by Spiekerman et al. (2017) in their "Anti-Transhumanist Manifesto" that directly opposes a vision of the human as merely information objects no different than other information objects (non-human agents) which they among others describe as "an expression of the desire to control through calculation" (p. 2). In this way, the human-centric approach can therefore also be depicted as a more general concern with the agency of humans in the control structures of their technical environments (Deleuze, 1992)

Building on this conception of a human-centric approach, I in the following purposefully make a distinction between the "human" and "non-human agent". In other words, AI can be defined as complex socio-technical data systems invested with interests and providing agency to different stakeholder group interests in society, but at the same time the very conception of the autonomous agency of AI can be considered an interest to be harnessed. As such, a human-centric distribution of power may be operationalized with design components of AI that enable the agency and critical reflection of human beings in semi-autonomous systems, and we may therefore also argue that it is this very human component of the AI system's design and use that should be the focus of a human-centric ethical governance approach to AI.

Data Interests by Design

Technological design is a dynamic process that is simultaneously shaped by society and is society-shaping (Hughes, 1987, p. 51). It embodies a range of diverse factors technological, social, political, economic, as well as the individual professional skills and prejudices of engineers that are all “thrown into the melting pot whenever an artifact is designed or built” (Bijker & Law, 1997, p. 3). Nevertheless, as Bijker and Law (1997) put it, technologies do not represent their own inner logic; they are shaped, sometimes even “pressed”, into a certain form that “might have been otherwise” (p. 3). From this perspective, we may consider the embedded data interests in the very technology design as a component in a network of factors that to a certain extend can be “ethically” guided. This very focus on design as one component of ethical governance may also be referred to as an “ethics by design” approach that aims to develop methods and tools for designing ethical behavior in autonomous agents to warrant that they behave within “given moral bounds” (Dignum et al., 2018, p. 61).

The idea that human values intentionally can be designed into a computer technology was originally formulated by Batya Friedmann and partners in the 1990s and has since then been further explored in the Value Sensitive Design (VSD) applied ethics framework (Friedman, 1996; Friedman et al., 2006; Friedman & Nissenbaum, 1995, 1997). In VSD, the embedded values of a technology are addressed as ethical dilemmas or moral problems to solve in the very design of computer technologies. A data interest may, in this context, be described as an intention or a motive that is transformed into specific properties of a data technology that arranges data in ways that support the agency of certain interests in the data that is stored, processed and analyzed by the AI technology. An applied ethics VSD approach would here aim to solve conflicts of data interests in the very design of a technology. For example, one could consider the individual’s interest in “privacy” as a value that is adversely affected by a specific design of a data-intensive technology, and accordingly suggest an alternative design in which privacy is deliberately designed into that technology (for example with a “privacy-by-design” [Cavoukian, 2009] approach). In VSD, interests are correlated with values that are held by stakeholders and that can be affected adversely or supported by a technology’s design. Thus, the aim is to develop analytical frameworks and methodologies to rebalance the distribution of interests and attempt at resolving conflicts between the needs and values held by these stakeholders in the design of a computer technology (Umbrello, 2019; Umbrello & De Bellis, 2018). These stakeholder analyses may also be extended to policy contexts in which technology design is negotiated. Steven Umbrello (2019, p. 7) for example identified specific values (data privacy, accessibility, responsibility, accountability, transparency, explainability, efficiency, consent, inclusivity, diversity, security, control) in the committee evidence reports of the UKs Select Committee on Artificial Intelligence tracing them directly to the different stakeholder groups involved in the committee (Academics, non-profits, governmental bodies and Industry/for profits) ranking their order of distribution).

Data Interests in Society

Data interests can be explicitly examined during the different design phases and the deployment of AI. Every day negotiations are taking place among different interests in society in very tangible digital data resources. These data interest negotiations we see represented in concrete deployments of AI technologies. They represent micro individual stakeholder objectives, values and needs, from developers to the users, institutional or business interests in data or they may even represent macro cultural sentiments or social structural requirements. Data interests may be aligned or compete, some are explicitly considered, but many are not considered at all in the developmental process or deployment of AI. An example is the "babysitter app" Predictim that uses language-processing algorithms and image-recognition AI software to analyze babysitter job seekers' social media posts to produce a personality report with a risk score (Harwell, 2018). At first glance, the app developers have an interest in bettering the AI functionality by enriching it with data from social media. From the users of the app's point of view, they have (as employers) an interest in the insights provided by the data analysis of the app. The job seekers on the other hand might have an interest in keeping their social media data private or in just reviewing the report based on their data, which is only accessible to employers. Facebook and Twitter also had an interest in the social media data processed by the app and therefore banned the app from their portals. Predictim's CEO said the company was not doing anything wrong: "Everyone looks people up on social media, they look people up on Google" (Lee, 2018). What he did not consider was a shift in the general societal interest in the data of AI; that is, an increasing concern with the social implications of big data technologies.

As a classic sociological concept, interests are considered major determinants of social action (Spillman & Strand, 2013, p. 86). Interests are sustained and amplified in social processes that determine the power dynamics in society. Recent studies from different scholarly fields have illustrated concrete examples of the social implications of AI, highlighting the power relations and interests at play in the development and societal adoption of AI. Data systems and mathematically designed algorithms are not impartial or objective representations of the world but are invested with the interests of the powerful—governments, public institutions and big data industries. They therefore trigger actions with ethical and social implications.

Examples of the social and ethical implications of AI and data intensive systems in general are manyfold. The mathematician Cathy O'Neil (2016) describes what she refers to as ”weapons of math destruction” (WMDs) that are deployed by private and state actors without question as neutral and objective systems replacing human decision-making and assessment with devastating consequences for citizens. A teacher loses her job due to a rigid machine-based performance assessment that does not take into account social contexts and human factors; a young person from a rough neighbourhood gets a visit from the police based on their use of a predictive crime tool. The legal scholar Frank Pasquale (2015) equally worries about the increasing use of automated processes to assess risks and allocate opportunities. These, he argues, are controlled by private companies that are also the most profitable and essential parts of the information economy. Another famous case study of the power relations at play in the adoption of AI systems is the "machine bias" study published by the news site Probublica. Here, the investigative journalist Julia Angwin (2016) together with a team of journalists and data scientists examined the private company Northpoint's COMPAS algorithm, which is used to perform risk assessments of defendants in the U.S. judicial system and assess the likelihood of recidivism after release. They found a bias against black defendants in the algorithm that had the tendency to designate black defendants as possible reoffenders twice as much as it did with white defendants. At the same time, it more often grouped white, rather than black, defendants as a low risk.

The Delegated Agency of AI

Langdon Winner (1980) describes technologies as "ways of building orders in the world", as active "structuring activities" by which "different people are differently situated and possess unequal degrees of power as well as unequal levels of awareness" (p. 127). Technologies are not neutral tools but have embedded politics, he argues. They are the locus of the distribution of societal powers shaped by human motives, and therefore they also embody "power and authority" (Winner, 1980, p. 131).

In a big data socio-technical environment, we may also see ourselves as locked in specific positions prescribed by the "politics" of different combinations of data sets that either grant us access or restrict access to insights and connections that can be transformed into lost or found opportunities**.** Bolukbasi et al. (2016), for example, examined how gender bias in data from news articles is reproduced and amplified in some of the most popular machine-learning methods for language processing that are used in online search engines. They found that due to existing bias in the training data of the model, words are organized in clusters of words such as architect, philosopher, financier and similar titles grouped together semantically as "extreme he" words, whilst words such as receptionist, housekeeper and nanny are grouped together as "extreme she" words (Bolukbasi et al., 2016, p. 2). An employer's online search for a candidate for a job position, might therefore very well present to her a list of information embedded with this type of pre-existing societal bias. This list represents the delegated moral agency of humans and society, which prescribes a very specific prioritization of the information she should look into. Whether she acts on the information she accesses based on this list or not, the design of it can still be argued to have embedded "politics" and "values" that are inscribed in the algorithm of the search engine. Thus, we may argue that choices in regard to the very data design of an AI system have potential implications for the agency of different data interests.

Deploying a perspective from economic theories on agency, we may here consider the informational relations between "agents" that manage the (social and/or economic) interests of "principals". To provide an example, an agent could be a real estate agent, and the principal could be the buyer and/or seller of a property whose primarily economic interests are negotiated in the information contained and shared in contracts and relationships. Applying this perspective in social theory, the fundamental idea is that we very often in society do not negotiate our own interests. We have allocated agents that negotiate on our behalf (Shapiro, 2005). That is, we delegate decision-making authority to agents that are to represent our interests. Thus, an agent could for example also be a nation state, and the principals could be the citizens whose agency is affected by the informational relationships they have with their government.

The concept of "informational asymmetries" between principals and agents and consequently trade-offs between interests is of core importance to a critical analysis of agency (Shapiro, 2005). A state's unaccounted-for mass surveillance of citizens constitutes an informational asymmetry; so too does a social media company's opaque harvesting and processing of the personal data of its users. Both forms of asymmetry have a direct effect on citizens' and individuals' agency.

The idea that our agency is defined by the informational relationships we have with different types of economic, social or political representatives that negotiate interests on our behalf is relevant to a data interest analysis of AI. That is, increasingly we delegate decision-making to AI technologies with built in trade-offs between different interests that might be correlated with or in conflict with our own interests. Hence, we may consider AI technologies as "moral agents", not with moral agency as such (Adam, 2008) and capable of their own of making moral decisions but as agents to which humans delegate the enforcement of the agency of different interests (Latour, 1992). AI technologies can in this way be described as agents representing our interests and acting on our behalf. They are nodes that manage informational relationships between different societal actors, distribute informational resources and allocate agency of interests.

To this end, the very design of the informational relationship (the data design) we have with our AI agents, the insight and access to data, constitutes the balancing and trade-offs between interests. Data design might embody "informational asymmetries", the implicit other of what in economic agency and game theory is referred to and aspired to as a desired state of "perfect information" in which all interests are served by having an equal amount of information to make rational decisions (Von Neumann & Morgenstern, 1953). Of course, technological development is always embedded with societal trade-offs and consequently ethical choice. Information is never perfect. It has a dimension of moral choice that entails trade-offs between interests, and it is this choice that requires an ethical governance framework.

Ethical Governance

If we consider AI data design as a type of agent with a particular organization of data interests that supports or represses their respective agencies, a core question to pose to the development of AI is if an alternative data design would balance out conflicts of interests in a human-centric framework to the benefit of the individual human being. The idea that we can design and govern technologies with an aim as such in mind, or an "ethics by design" approach, calls for an emphasis on the design of the moral qualities of the very technologies, that is, respectively their embedded "values" or their "politics". However, this also needs to be understood more generally in terms of distributed governance in which the "values sensitive design" of the very data design of AI is only one component. Thus, when we examine the ethical implications of AI-based systems, we need to look at the way in which they are distributed among active human and non-human actors and we need to understand them as components of complex social, cultural and political environments. This also means that we cannot just design a moral norm into a technology and in this way produce a “moral machine”, we also need to address the way in which ethical implications evolve in environments of distributed moral agencies between human and non-human actors (Latour, 1992, 2002) that are in constant negotiation with use, laws and standards, culture and the society. Social and ethical implications of AI systems are always the result of a combined network of actions and competences distributed between different agencies – the data design, the engineers, the users, norms, laws and culture etc. If we for example consider the gender bias of a “word embedding“ method for online search, it is not only a property of a particular AI design. The bias and discrimination result from a distribution of agency between the machine learning model (non-human actor) that is actively amplifying existing human bias in its training data, learning and evolving from Google news articles (society and culture) and personalized (the user data), prescribed by data design (the developer’s intentions), developed, accepted and enacted in society as an “objective” representation of information (culture), particular interpretations and implementations (or lack thereof) anti-discrimination or data protection legal frameworks (law), engineering standards and methods (scientific paradigms) and a range of other factors. This is why we need to think of a human-centric approach to the data interests of AI as not only a framework for the design of an AI technology, but moreover as a framework for the design of the “ethical governance” of AI in general.

Winfield and Jirotka (2018) present a case for “a more inclusive, transparent and agile form of governance for robotics and artificial intelligence (AI) in order to build and maintain public trust and to ensure that such systems are developed for the public benefit” (p. 1). Ethical governance, they argue, goes beyond just effective and good governance, but is “a set of processes, procedures, cultures and values designed to ensure the highest standards of behaviour” (Winfield & Jirotka, 2018, p. 2). They address the socio-technical nature of AI development and adoption in which the technical is intrinsically intertwined with the social and vice versa. Governing the development of robotics and AI with an ethical framework therefore requires a diverse set of approaches from those at the level of individual systems and application domains to those at an institutional level. Winfield and Jirotka are addressing the “ethical governance” instruments that companies need to develop and adopting AI ethically. But here I also want to propose that the “ethical governance” of AI concerns the applied components of AI ethics in the general governance of socio-technical systems with a distributed form of power in which data interests play a key role.

In what follows, I derive data interest themes from the “Ethics Guidelines for Trustworthy AI” (2019) developed by the AI HLEG appointed by the European Commission. It is important to note here that in isolation the guidelines do not ensure a human-centric AI development. That is; as stand-alone principles they are not the solution to an ethical problem. Winfield and Jirotka (2018) present a map that connects ethical principles with emerging standards and regulation, including a level of verification and validation. They also argue that “while there is no shortage of sound ethical principles in robotics and AI, there is little evidence that those principles have yet translated into practice, i.e. effective and transparent ethical governance.” (p. 9) That is; ethics guidelines, such as the AI HLEGs do not “govern”, but they may inspire, guide and even set in motion political, economic and educational processes that foster an ethical “design” of the big data age, which means everything from the introduction of new laws, the implementation of policies and practices in organisations and companies, development of new engineering standards, to awareness campaigns among citizens and educational initiatives.

Ethics Guidelines for Trustworthy AI

The High-Level Expert Group developing the guidelines consisted of 52 members representing stakeholder interests from industry, civil society and academia. Although set up in a policy context, the group was not a policy body per se. The European Commission sets up multi-stakeholder high-level expert groups to inform their policies in different areas. These are independent, and the Commission does not participate directly in the groups’ work, nor do they necessarily use the work of these groups directly in policymaking. The ethics guidelines were developed over a one-year period and included a public consultation process for the first draft in which comments were received and incorporated in the text. Although this very multi-stakeholder process by itself would be a relevant focus for a data interest analysis, I in this article focus only on the very text of the guidelines in order to illustrate a negotiation of data interests with a human-centric applied ethics approach.

The guidelines specifically aim to provide AI practitioners with a methodology for incorporating a human-centric framework in the design of AI and as follows also for the governance and management of the data of AI. Three core components of Trustworthy AI are proposed in the guidelines. AI should be lawful—respect laws and regulations, ethical—respect ethical principles and values, and robust—from a technical perspective and with consideration of its social environment. These three components are the essential framework within which the concept of Trustworthy AI is interpreted. First of all, they refer to an existing legal framework; that is, the guidelines are not an alternative to or an interpretation of law. The first component, legal compliance, is therefore assumed as the basis of Trustworthy AI. The second ”ethical” component is expanded in four ethical foundational principles: respect for human autonomy, prevention of harm, fairness and explicability. The third component is meant to ensure the robustness of the AI system to guarantee the safe, secure and reliable behaviour of the system to prevent unintended adverse impacts. The guidelines also present seven key requirements that AI systems should meet in order to be deemed trustworthy: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental accountability and auditability. An assessment list is here included, aimed at operationalizing these key requirements when developing, deploying or using AI systems. This list was revised in 2019–2020 in a piloting phase to include input from companies and institutions developing AI in different sectors.

The Five Data Interests and an Example

In the following, illustrate a human-centric approach to the data interests of AI development based on an analysis of the data metaphors of the AI HLEG ethics guidelines[i]. I suggest five clusters of themes and sets of questions that might help forward an exploration of data interests in AI design and development: data as resource, data as power, data as regulator, data as vision and data as risk. I then exemplify a human-centric approach to data interests in a case in which an AI design takes explicitly point of departure in the data interest of the individual human being.

This analysis is not intended as a comprehensive framework the implementation of a human-centric approach to data interests by design, nor does it represent a complete analysis of the data interests represented in the ethics guidelines. Importantly this is only a values-based interpretative framework that complements but does not by any means replace the implementation of legal frameworks. What I mean to do here is to present a human-centric “compass” that may help guide an applied ethics by design approach to data interests in AI development and governance.

I. Data as Resource

Who or what provides the data resource? Who or what has an interest in the data resource? How is the data resource distributed and how does the human being benefit?

The “data as resource” cluster concerns the very distribution of data resources among involved interests in the data design. In the ethics guidelines, data is considered a resource, metaphorically separated from that which it represents (a person or an artifact). Data can be “provided”, “accessed”, “gathered”, “labelled”, “extracted”, “used”, “processed”, “collected”, “acquired” and “put” into a system in a “structured” or “unstructured” (words used throughout the guidelines to describe the data of AI). A resource is a corporeal and spatially delineated thing, something we can be in possession of, place in containers or create boundaries around, store and process, and it is something we can be with or without. The resource metaphor is common in public discourse on big data (Puschmann & Burgess, 2014). Sarah M. Watson (n.d.) refers to the dominant metaphors for personal data in public discourse as “industrial”, as if it were a “natural resource” to be handled by “large-scale industrial processes”. Mayer-Schonberger & Cukier (2013) depict data as the raw material of a big data evolution of the industrial age. But the “data as resource” metaphor actually denotes two different things. One type of resource in society is indeed the raw material of industrial processes that is processed and turned into products. Another type of resource is the kind that makes us stronger as individuals. Being resourceful also means that you as an individual have the capacity, and physical, psychological and social means. The first type of resource mentioned here is tangible and material, the other being social and psychological. But both are what Lakoff and Johnsson (1980, p. 25) would refer to as “container” metaphors, entities with boundaries that we can handle and reason about. Subsequently, data as a resource can be protected and governed in very tangible ways, and in each case micro and macro stakeholder interests in these governance and resource handling frameworks are involved. Very broadly speaking, industries have an interest in the raw material of their data-based business models: political players have an interest in providing rich data infrastructures for AI innovation to compete in a global market; the engineer has an interest in volumes of data to train and improve the AI system; individuals have an interest in protecting their personal data resourcefulness or even enhancing their data resources by creating their own data repositories and accordingly benefitting directly from these (as the “data trust” or “personal data store movement” represents). Obviously, treating the psychological and social resources of individuals as material resources in an industrial production line represents a conflict of data interests. But, in addition, informational asymmetries also create very tangible social and economic gaps between the data rich and the data poor, which is a conflict of interest on a more general structural level of society. In the guidelines, there is a reference to data as the raw material resource of AI technologies (“its evolution over the last several years has been facilitated by the availability of enormous amounts of digital data" p. 35). However, data and the governance of this is first and foremost associated with privacy and the protection of the individual human being (the third of the guidelines' requirements, p. 17). Closely connected with the ethical principle "prevention of harm", the protection of data as a social and psychological resource of human beings is in the guidelines described as "protection of human dignity as well as mental and physical integrity" (p. 12). This includes particular attention to socially vulnerable groups or individuals. Proper governance of the personal data resource is described as the primary means to "allow individuals to trust the data gathering process" (p. 17) and consequently forms a foundation the development of "Trustworthy AI". To provide an example, an AI developer will make use of different resources of data for the design of an AI system. These data resources they have to either process on their own computers or they can use the more powerful cloud-based AI platforms of for example Google or Amazon. However, by doing this they also have to share their data resources with these actors. If prioritizing individuals' data interests, when dealing with for example personal data, they have to trust that these companies do the same. This entails a data interest design choice. Moreover, different actors may have interests in the data resource. For example, if the developer is creating an AI system for assessing the performance of employees, the employee might have an interest in creating a bigger data resource that can assess minute details of the employees work day. The employer might even want to collect data from outside the work place from the employees' social media presence for the AI system to predict potential risks to the work place, such as for example job searches. The employees on the other hand have an interest in keeping the data resource to a specific limit and a union representing these employees might also want to get involved to safeguard the data interest of the people they represent.

II. Data as Power

Who is empowered or disempowered by the data access and processing? Does the data design support the human being's power and data agency? How are conflicts between different data interests resolved to the benefit of the human being?

The second cluster, "data as power", is closely connected with the first cluster, "data as resource", as the distribution of resources also constitutes a distribution of power. A society is based on balances of power between social groups, states, companies and citizens. A democratic society, for example, represents one type of power structure in which the governing powers are always balanced against the level of citizen power. Distribution of

data/information amounts to the distribution of power in society. David Lyon envisions an "ethics of Big Data practices" (2014, p. 10) to renegotiate an unequal distribution of power embedded in technological big data systems. In the guidelines, changing power dynamics are addressed in light of the information asymmetry between individuals and the institutions and companies that collect and process data in digital networks, where AI systems are the interlocutor (the agents) of this type of informational power: "Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens." (p. 12). Access to or possession of data is here associated with dominance of some societal groups/and or institutions, businesses over others, but also with the very function of democratic/and or "fair" processes, where access to information and explanation of data processing ("the principle of explicability", p.13) is the basis of "accountability" and "fairness" ("Without such information, a decision as such cannot be duly contested.", p. 13). Data design solutions suggested are here associated with the concept of "human agency" that supports "informed autonomous decisions regarding AI systems" or "informed choices in accordance with their goals" (p. 16).
For example, access to the data and data processes of a system can provide an investigative journalist or researcher with the power to challenge an AI system. The researcher has an interest in the data of the system that can be inhibited by a company's interest in keeping algorithms and data design proprietary. In connection with the Machine Bias study of the Compass software, the researchers did not have access to the calculations used for defendents risk scores. At one point they received the basics of the future-crime formula from the company behind, Northpointe, but it never shared the specific calculations, which it said are proprietary (Angwin et al, 2016).

III. Data as Regulator

Which interests will the implementation of law serve and prioritize? Does the data design of the AI system truly enhance the values of the law, or does it just barely comply? Are there conflicts between big data interests and the embedded values to protection of the individual human being of a legal framework such as the GDPR? How are they resolved in the design?

The third cluster, “data as regulator”, represents the legal enforcement of the power balancing of data interests. In an ideal constellation, law and technology design supplement each other. Technology design is a type of “regulator” that either protects or inhibits legal values. “Code is law”, as Lawrence Lessig (2006) formulates it quoting Joel R. Reidenberg’s concept “Lex Informatica” that links technological design choices and rule-making with governance in general and understanding the two as inseparable, with the first enforcing the other (Reidenberg, 1997, p. 555). Although it is emphasized that the guidelines do not replace existing laws, and therefore the first component of Trustworthy AI, “lawful AI”, is only a reiteration of compliance with legal frameworks, it can be argued that the guidelines still treat the “Lex Informatica” of lawful AI: “The guidelines do not explicitly deal with the first component of Trustworthy AI (lawful AI), but instead aim to offer guidance on fostering and securing the second and third components (ethical and robust AI)”(p. 6). Throughout the text, the design of AI is associated with the implementation of legal principles. The data design of an AI technology is here described as essential to the auditability and accountability of an AI system (p. 19) that ensures that it can be assessed and evaluated as well as be able to implement the principle of explicability (“Without such information, a decision cannot be duly contested.” p. 13). In this way, the “data as regulator” metaphor emphasizes the role of a particular data design in the legal implementation and realization of law in society (“AI systems can equally enable and hamper fundamental rights.” p. 15). That is also to say that even though compliance with the European General Data Protection Regulation (GDPR) is considered in the development of an AI technology, it may or may not actually realize the legal principles of this law in its very data design. In a micro-engineering context, the design challenges of a data-intensive technology such as AI are, for example, the development of a data design implementing principles such as data minimization, privacy (data protection) by design or the right to (information) an explanation. Different stakeholder interests in the data design may also be in conflict with legal frameworks such as the Fundamental Rights Framework in which the data interests of individuals is a primary value. For example, non-democratic state actors have an interest in data for social control purposes; business interests in tracking and collecting data to enhance their data-based business models without consideration of individuals’ rights, or similarly scientific interests in improving research with big data analytics without these considerations. The ethics guidelines mention some of the greatest legal challenges of AI in a European fundamental rights framework—mass surveillance, discrimination, the undermining of democratic processes—and propose alternative design options embodied in the concept of Trustworthy AI.

IV. Data as Eyes

Can the human beings involved in the design and implementation "see" the data processes and their implications? What does the data design see (the training data) and then perceive (how is it instructed to act on training data)?

The fourth cluster, "data as eyes", represents the very agency in the data design of data interests that here may be equated with vision (what we are able to see or not, and how we see it) in digital data systems. In an ideal constellation, vision is an effortless extension of our eyes. However, in a digital data-based environment, the instruments (our digital eyes) that we use to see and perceive our environment are quite literally extended into data design that constitutes a management of what we can see.

As previously discussed, data design also functions as a moral agent in that it prescribes and manages our active engagement with the information it handles as well as the digital information infrastructure in which it exists. In the ethics guidelines, eyes and vision are embodied in data systems. While data is described as the technology's sensory system on which it develops its mode of action ("perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal", p. 36), data is also the "eyes" of the individual. Data will for example "yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used…" (p. 18). The concept of "transparency" is used in the most literal sense to denote our ability to see the data processes that should be documented and made traceable. "Black box" algorithms' (Pasquale, 2015) processing of data may in this context blur our vision or make us blind to the reasoning of an AI system. The management of visibility, that is, the very architecture of visibility of emerging technological environments can be said to constitute a mode of social organization and distribution of powers (Brighenti, 2010, Flyverbom, 2019). What is made visible, what remains invisible and importantly who is empowered to see through the social organization of visibilities directly influences the agency of interests in society. The eyes as the data design of an AI technology do exactly that. To provide an example, we might create an AI system to analyze the data of people on social welfare benefits. This may have a dashboard for the public institution's social workers, which provides general data statistics, fraud detection and risk scores on individuals. The dashboard is our eyes within the AI system. In this case, only the social worker's data interest has eyes and so does the public institution's interest in controlling public resources and optimizing work processes. But we could also think of other types of data design in which the people on social welfare have eyes through data access and hence agency to, for example, add and correct flawed data or personalize the services provided to them.

V. Data as Risk

To whom or to what could the data design be a risk? Who or what has an interest in preventing and managing the identified risks? How are conflicts and/or alignments between identified risks to the human being and interests in managing the risks resolved to protect the human being from risks?

In the fifth and last cluster, "data as risk", data interests are correlated with the act of assessing the risks of data design. The risks of the development and deployment of AI technologies are addressed generally in the guidelines as the potential negative impacts of AI systems on democracy and civil rights, on the economy, and on markets and the environment. The risks are the harms of AI that should be anticipated, prevented and managed (p. 2). In particular, in the third core component ("robust AI"), a preventative risk-based approach to AI is emphasized: "Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner, such that they reliably behave as intended while minimising unintentional and unexpected harm" (p. 16). Risks are not just a particular concern of these guidelines, but of contemporary politics, business conduct and public discourse in general. Ulrich Beck described this societal preoccupation with risk prevention and management in his seminal book *The Risk Society* (1993) as an uncertainty produced in the industrial society, and as the result of a modernization process, in which unpredictable outcomes are emerging and accumulating (Beck, 1992). "Risks are not 'real', they are 'becoming real'", Beck (2014, p. 81) later proclaimed when describing a development of concern with "world risks". They take the shape of "the anticipation of catastrophe" (Beck, 2014, p. 81) and their management as an "anticipation of further attacks, inflation, new markets, wars or the restriction of civil liberties" (p. 81). Importantly, the depiction of a risk "…presupposes *human decisions*, human made futures (probability, technology, modernization)…" (Beck, 2014, p. 81)

In the AI HLEG guidelines, the data of AI is described as laden with potential risks that should be prevented and managed. As described in the four other metaphorical clusters, data is generally associated with the management of risks to resources in society, to democracy, to the rule of law and to agency through visibility. But in this last metaphorical cluster of the guidelines, data is a risk per se that must be anticipated and managed. Criminals can for example "attack" the data of an AI system and "poison" it (p. 16), data can "leak", data can be "corrupted by malicious intention or by exposure to unexpected situations" (p 16), and it can pose a risk to the environment ("Did you establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (for example, the type of energy used by the data centres)?", p. 30). Data by itself also has risky properties; for example, words such as "malicious" or "adversarial" are also used in the guidelines to describe data. If we continue with the notion that risks are not "real" per se but based on our own predictions about possible future scenarios, then we may also assume that their proposed management and prevention is the product of interests and motives. On the AI development side, an AI engineer training an AI technology will have an interest in the risks posed to the quality and accuracy of the training data, while a data protection officer will consider risks posed to identified individuals with a data protection impact assessment. On the AI deployment and adoption side, an individual will also have a "data-as-risk" interest in keeping their personal data safe from unauthorized access, while an anti-terror intelligence officer's risk scenario involves the detection of terrorist activities and might therefore consider the end-to-end encryption of a data design a risky design choice. The consideration of different "data-as-risk" scenarios guided by the different interests involved in these will direct disparate design choices that might be aligned but might also be in conflict.

An Example: P7006 The Personal Data AI Agent

The IEEE P7006 standard for Personal Data Artificial Intelligent (AI) Agents (PDAIAs)[ii] which is one of the IEEEs P7000s series of ethically aligned design standards for autonomous and intelligent systems. The standard was work in progress at the time of writing this article, but the PAR and work in progress (wp) principles (P7006, 2020) can be used to illustrate how reflections on the operationalization of some of the thematic clusters of a human-centric framework for data interests can be translated into a technical standard for the design of an AI technology.

Personal AI assistants are software programs that interact with individuals answering questions and performing different tasks for them. They are based on various AI components such as voice recognition and machine learning to act on behalf of and respond to individual users in a personalized manner. Personal AI assistants have been popularized by leading commercial actors as integrated components of mobiles, tablets or speakers with for example Amazon's Alexa, Apple's Siri, Google Now and Microsoft's Cortana (Bonneau, Probst & Lefebvre, 2018). As virtual agents integrated in individuals' private sphere at home and out that recognize voices, store and process users' personal data, several critiques and concerns have raised in the public debate. In particular privacy concerns and the prioritization of the power and interests of the commercial actors and other actors versus that of individuals (Lynsky, 2019, Nelius, 2020, Chung, Iorga, Voas & Lee, 2017, Maedche, A., Legner, C., Benlian, A. et al., 2019).

Against this background, the P7006 standard's scope is explicitly defined in its Project Authorization Request (PAR) as one that prioritizes the data interest of the individual human being: "This standard describes the technical elements required to create and grant access to a personalized Artificial Intelligence (AI) that will comprise inputs, learning, ethics, rules and values controlled by individuals." (P7006, 2017) It is furthermore in the PAR envisioned as an agent that first and foremost manages the data interests of individuals enabling them to "safely organize and share their personal information at a machine-readable level and enable a personalized AI to act as a proxy for machine-to-machine decisions (P7006, 2017). The PDAIA is trained and evolves with the data of an individual. As such the main data resource is that of the individual. The data interests in this data are described in the scope and purpose of the standard: "The primary stakeholders for this standard are the individuals whose information is utilized for any data interaction, but also includes the academics, engineers, programmers, marketers or technologists of any kind wishing to utilize said data" (P7006, 2017). However, the data interest benefitted most profoundly by the data resources is that of the individual in that the AI data design is meant to serve as an agent for the individual that "negotiates their individual rights and agency in a system of shared social norms, ethics and human rights that also foresee and helps the individual mitigate ethical implications of data processing" (P7006, 2017). This prioritization of the individual's interest in data is sought operationalized throughout the wp principles, for example in one principle that is meant to ensure that the data resource is maintained by the individual human. The interest of the individual human being is moreover prioritized in the power distribution of the data design. For example, the same principle aims to ensure that the power over databased decisions remains with the individual when stating that the PDAIA will remain semi-autonomous and never fully independent of the individual while making decisions. As such, another princple concerns “Interruptible & Retractable” decision-making and among others includes the incorporation of a “Kill Switch” at root-level design that the individual user can use to halt and or reverse the actions of an agent. The data design’s management of visibility is moreover specifically meant to provide the individual human being with vision of the data processes and thereby agency. The first principle for example entails “human-interpretable ”transparency”” which is further delineated with the provision: ”“Black box” processes must be able to be tracked in a way that creates a human-readable chain of custody for all decisions”. The vision of the individual human being is also extended to include information regarding third-parties that have influenced a decision. Data as a type of regulator is considered with specific consideration of “privacy” design rather than data protection and security only for example in one principle that considers a “Privacy-oriented dynamic UI/UX. A user can easily toggle Agent privacy settings to enable updating system settings to facilitate differential privacy depending on the occasion.” Finally, data is primarily considered a risk to the human being and accordingly one principle considers the safety of the user based on user-centered agency, protection against ”physical, empotional or cognitive harm” and safeguards from revealing compromising information from law enforcement.

Conclusion

Over a very short period of time there has been an acceleration in the integration of data-intensive technologies such as AI —a complex and advanced machinery designed to act on data, predict behaviour and trajectories, and point lives, economies, politics, society in specific directions— in all areas of our lives. AI technologies constitute a distribution of informational resources in society, and they therefore have a direct influence on the balance of agency among societal actors.

In this article, I have made a case for a data interest analysis of AI technologies to identify data interests in their data design and to explore how these interests are empowered or disempowered by design. I have also put forward a specific applied ethics approach to data interests based on the human-centric ethical governance framework presented in the EU High-Level Group on AI’s Ethics Guidelines for Trustworthy AI.

AI systems are not impartial or objective representations of the world. They reinforce societal power relations and are invested with the interests of the powerful, and they therefore trigger actions with ethical and social implications. Big data industries have an interest in harvesting and accumulating data, improving their machine-learning models or earning a profit on personalized marketing. Public institutions have an interest in streamlining and controlling public administration. Governments have an interest in controlling, viewing and predicting risks to their nations from data. Global industry and intergovernmental bodies have an interest in gaining a competitive advantage with data. Obviously, interests are not as uniform as this, nor are they simple determinants of action in society. But as we see time and again when ethical implications of AI systems materialize into social implications, data interests are present in the design of AI and are provided with agency or no agency. The result is increasing informational asymmetries among people, such as between civil servants, developers or people from minority groups and societal stakeholders, such as industries, governments and citizens.

An applied ethics approach to AI systems highlights the social power relations and interests at play in the development and societal adoption of AI. Narrowing down an approach as such to the data interests in AI allows us to consider the built-in trade-offs between these different interests, and it opens up the possibility of an exploration of alternative data designs that may solve conflicts of interests in a way that benefits the human being. Here, fostering an awareness of the primacy of the human interest over commercial, state, and even scientific interests in the development of big data AI systems is one such awareness that needs a strong position in the AI ethics debate.

[i] To investigate the themes, I used a metaphorical analysis (Lakoff & Johnsson, 1980). I first went through the text to detect metaphors concerning data connecting them with common metaphors concerning data in public discourse. Then I organized the text according to the core metaphors I found with a focus on the data interests they represented, their conflicts and finally how they were sought balanced in a human-centric framework.

[ii] I am the vice chair of the standard working group which is chaired by Ken Wallace. The PAR and work in progress principles are publicly available on the P7006 working group website. The work in progress draft principles are created by the members of the P7006 working group with the objective to guide the more detailed work on the standard. They are not complete, subject to negotiation and revision, and should therefore also be viewed as such. The specific draft principles that I refer to here were substantially revised by working group member Mathana Stender in June 2020. Although in draft, I have used them here to illustrate an attempt to operationalise an approach to data design that prioritizes the (individual) human being's interest.

References

AI HLEG/High-Level Expert Group on Artificial Intelligence (AI HLEG), 2019 "Ethics Guidelines for Trustworthy AI". April, 2019. https://ec.europa.eu/digital-single-market/en/news/ethics- guidelines-trustworthy-ai

AI HLEG/High-Level Expert Group on Artificial Intelligence (AIHLEG B), 2019. "Policy and investment recommendations for trustworthy AI. https://ec.europa.eu/digital-singlemarket/en/news/policyand-investment-recommendations-trustworthy-artificial-intelligence

Adam, A., 2008 "Ethics for things". *Ethics and Information Technology*, *10*(2–3), 149–154. https://doi.org/10.1007/s10676-008-9169-3

Alpaydin, E., 2016 *Machine learning the new AI*. Cambridge, MA: MIT Press.

Angwin, J., Larson, J., Mattu, S., Kirchner, L., May 23, 2016, "Machine Bias", Probublica https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Beck U., 2014 "Incalculable Futures: World Risk Society and Its Social and Political Implications". In: Beck U. (eds) *Ulrich Beck*. SpringerBriefs on Pioneers in Science and Practice, vol 18. Springer, Cham

Beck, U., 1993 *Risk Society. Towards a New Modernity*. London: SAGE Publications

Bijker, E. W., & Law, J., 1997 *Shaping technology/building society studies in sociotechnical change*. MIT Press.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T., 2016 "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings" 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain

Bonneau, V., Probst, L., Lefebvre, V., 2018 "The Rise of Virtual Personal Assistants"; Digital Transformation Monitor, European Commission.

Brighenti, A. M., 2010 New media and networked visibilities. In A. M. Brighenti (Ed.), *Visibility in social theory and social research* (pp. 91–108). Palgrave Macmillan. https://doi.org/10.1057/9780230282056_4

Cavoukian, A., 2009 "Privacy by design. The 7 foundational principles". Canada: Information & Privacy Commissioner. https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Génova, G., Kließ, M., Lopez-Sanchez, M., Micalizio, R., Pavón, J., Slavkovik, M., Smakman, M., van Steenbergen, M., Tedeschi, S., van der Torre, L., Villata, S., de Wildt, T., & Haim, G., 2018. "Ethics by Design: Necessity or Curse?" In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society—AIES '18* (pp. 60–66). ACM Press. https://doi.org/10.1145/3278721.3278745

Deleuze, G., 1992 "Postscript on the societies of control" October, 59, p. 3-7.

Elish, M. C., & Boyd, D., 2018 "Situating methods in the magic of big data and artificial intelligence". *Communication Monographs*, *85*(1), 57–80. doi: 10.1080/03637751.2017.1375130

Eubanks, Virginia, 2018 *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. New York, NY: St. Martin's Press

European Parliament, "A comprehensive European industrial policy on artificial intelligence and robotics", 12.2.2019, Strasbourg. https://www.europarl.europa.eu/doceo/document/TA-8-2019-0081_EN.html

Fjeld, J., Hilligoss, H., Achten, N., Daniel, M. L., Feldman, J., & Kagay, S., 2019 "Principled artificial intelligence: A map of ethical and rights-based approaches" https://ai-hr.cyber.harvard.edu/primp-viz.html

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pa
gallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People white paper: "Twenty
recommendations for an ethical framework for a good AI society". *Minds and machines*.

Flyverbom, Mikkel, 2019 *The Digital Prism: Transparency and Managed Visibilities in a Datafied World*, Cambridge University Press.

Friedman, B., 1996 "Value-sensitive design". *ACM Interactions*, *3*(6), 17–23. https://vsdesign.org/publications/pdf/friedman96valuesensitivedesign.pdf

Friedman, B., Kahn, P. H., Jr., & Borning, A., 2006 "Value sensitive design and information systems". In P. Zhang and D. Galletta (Eds.), *Human-computer interaction in management information systems: Foundations* (pp. 348–372). M.E. Sharpe. https://vsdesign.org/publications/pdf/non-scan-vsd-and-information-systems.pdf

Friedman, B., & Nissenbaum, H., 1995 "Minimizing bias in computer systems". In *Conference companion of*
CHI 1995 conference on human factors in computing systems (p. 444). ACM
Press. https://vsdesign.org/publications/pdf/friedman95minimizebias.pdf

Friedman, B., & Nissenbaum, H., 1996 "Bias in computer systems". *ACM Transactions on Information Systems, 14*(3), 330–347. https://vsdesign.org/publications/pdf/64_friedman.pdf

Friedman, B., & Nissenbaum, H., 1997 "Software agents and user autonomy". In *Proceedings of first international conference on autonomous agents* (pp. 466–469). ACM Press.
https://vsdesign.org/publications/pdf/friedman97softwareagents.pdf

GDPR/Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Harwell, D., 2018, November 23, in press "Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude". *Washington Post*. https://www.washingtonpost.com/technology/2018/11/16/wanted-perfect-babysitter-must-pass-ai-scan-respect-attitude/

Hasselbalch, Gry (2019): "Making Sense of Data Ethics – The Powers Behind The Data Ethics Debate in European Policymaking", Internet Policy Review, 8(2).

Hasselbalch, Gry (forthcoming): Culture by Design – A Data Interest Analysis of the European Policy Agenda, FirstMonday

Hughes, T. P., 1987 “The evolution of large technological systems”. In W. E. Bijker, T. P. Hughes, & T. Pinch (Eds.), *The social construction of technological systems* (pp. 51–82). MIT Press.

Chung, H., Iorga, M., Voas, J., Lee, S., 2017 “Alexa, Can I Trust You?”, Computer, 50(9):100-104 10.1109/MC.2017.3571053

P7006, 2017 “Standard for Personal Data Artificial Intelligence (AI) Agent” https://development.standards.ieee.org/myproject-web/public/view.html#pardetail/6097

P7006, June 2020 “Work in Progress Principles for P7006” https://sagroups.ieee.org/7006

Jobin, A., Ienca, M., & Vayena, E., 2019 “The global landscape of AI ethics guidelines”. *Nat Mach In tell, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Lakoff, G., & Johnson, M., 1980 *Metaphors we live by*. The University of Chicago Press.

Latour, B., 1992 “Where are the missing masses? The sociology of a few mundane artifacts.” In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 225–258). MIT Press. http://www.bruno-latour.fr/sites/default/files/50-MISSING-MASSES-GB.pdf

Latour, B., 2002 (Translated by Venn, C.) “Morality and technology”. *Theory, Culture & Society*, *19*(5–6), 247–260. https://doi.org/10.1177/026327602761899246

Lee, D., 2018, 27 November, in press. “Predictim babysitter app: Facebook and Twitter take action”. *BBC News*. https://www.bbc.com/news/technology-46354276

Lehr, D., & Ohm, P., 2017 “Playing with the data: What legal scholars should learn about machine learning” *UCDL Review, 51*, 653–717.

Lessig, L., 2006 *Code version 2.0*. Basic Books.

Lyon, D., 2010 “Liquid surveillance: The contribution of Zygmunt Bauman to surveillance studies”. *International Political Sociology*, *4*(4), 325–338. Doi:10.1111/j.1749-5687.2010.00109.x

Lynsky, D., 2019, 9th Ocetober, in press “’Alexa, are you invading my privacy?' – the dark side of our voice assistants”, The Guardian.

Mayer-Schonberger, V., & Cukier, K., 2013 *Big data: A revolution that will transform how we live, work and think.* John Murray.

Maedche, A., Legner, C., Benlian, A. *et al.,* 2019 ”AI-Based Digital Assistants”. *Bus Inf Syst Eng* 61, 535–544

Nelius, J., 2020, September 4, in press ”Amazon’s Alexa for Landlords Is a Privacy Nightmare Waiting to Happen”, Gizmodo.

Noble, Safiya Umoja, 2018 *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press

O’Neil, C.,2016 *Weapons of math destruction*. Penguin Books.

OECD, 2019 “Principles on Artificial Intelligence” in OECD Council Recommendation on Artificial Intelligence, OECD/LEGAL/0449, 22.05.2019 https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Pasquale, F., 2015 *The black box society – The secret algorithms that control money and information*. Harvard University Press.

Reidenberg, J. R., 1997 "Lex Informatica: The formulation of information policy rules through technology".

Texas Law Review, *43*. 553-593

Searle, J., 1980 ”Minds, brains, and programs”. *Behav*ioral and Brain Sciences, *3*(3), 417-424.

Shapiro, S. P. (2005). Agency theory. *Annual Review of Sociology*, *31*(1), 263–284. https://doi.org/10.1146/annurev.soc.31.041304.122159

Spillman, L., & Strand, M., 2013 "Interest-oriented action". *Annual Review of Sociology*, *39*(1), 85–104.
https://doi.org/10.1146/annurev-soc-081309-150019

Spiekermann, S., Hampson P., Ess, C. M., Hoff, J., Coeckelbergh, M., & Franckis, G., 2017 ”The Ghost of
Transhumanism & the Sentience of Existence" Retrieved from The Privacy Surgeon
http://privacysurgeon.org/blog/wp-content/uploads/2017/07/Human-manifesto_26_short-1.pdf

Umbrello, S.; De Bellis, A.F. "A Value-Sensitive Design Approach to Intelligent Agents" In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; CRC Press: Boca Raton, FL, USA, 2018; pp. 395–410.

Umbrello, S., 2019 "Beneficial Artificial Intelligence Coordination" by Means of a Value Sensitive Design Approach. *Big Data and Cognitive Computing*, *3*(1), 5. MDPI AG. http://dx.doi.org/10.3390/bdcc3010005

Von Neumann, J., & Morgenstern, O., 1953 *Theory of games and economic behaviour.* Princeton University
Press.

Watson, S. M. (n.d.). "Data is the new ...." *Dis Magazine*. http://dismagazine.com/blog/73298/sara-m-watson-metaphors-of-big-data/

Winfield, A. F. T., & Jirotka, M., 2018 "Ethical governance is essential to building trust in robotics and artificial intelligence systems" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180085. https://doi.org/10.1098/rsta.2018.0085

Winner, L., 1980 "Do artifacts have politics?" *Daedalus*, *109*(1), 121–136.