Gry Hasselbalch, Draft 27th September, 2019.

The Data Ethics of Artificial Intelligence:

A Data Interest Analysis

Abstract:

This article presents a framework for a data interest analysis of artificial intelligence technologies (AI) that can be used to explore how different interests in data are empowered by design and to make reflective choices correspondingly. While common public discourses often present uncontrollable AI future scenarios, this data interest analysis looks at AI as complex data processes formed by societal interests and maintains that our intelligent data intensive technological environment can be actively shaped and directed. The analytical framework is based on an exploration of data interest metaphors in the European Commission's High Level Group on AI's Ethics Guidelines for Trustworthy AI published 8th of April 2019.

Introduction:

"The numerical language of control is made of codes that mark access to information or reject it" (Deleuze, 1992, p. 5).

Data flows. Data transforms. Data changes hands, bodies and containers. Made possible by the endless amounts of the data that we generate in society, artificial intelligence technologies (AI) is the next frontier in the age of big data flows. Developed and adopted to contain and make sense of large amounts of data, predict patterns, analyze risks and act on that knowledge with decisions in politics, culture, industries, and on life trajectories. In this article I propose a conceptual framework for a data interest analysis of AI. It can be used to explore how different interests in data are empowered by design and to make choices in AI design, policy and education accordingly. As the "raw material" of this age of big data (Mayer-Schonberger & Cukier, 2013) the data of AI is heavily interest invested. Negotiations between different interests in data are taking place every day concerning very tangible digital data resources, data acquisition, data access, data analytics, data ownership and the attached values. Legislators are redesigning data protection legal frameworks, while at the same time seeking to harness the power of data monopolies with antitrust and competition legislation, and states and companies alike are carving their space in this age of big data with aggressive data harvesting strategies. Meanwhile citizens are toiling to understand how their interests are preserved in big data technological environments. We need to understand how these interests in data join, appear and/or disperse in the development of AI as an at the same time symptom of and condition for the distribution of agency and powers in society.

The point of departure for this data interest analysis is a "data ethics of power" which can be summarised as the general cross-disciplinary study of the distribution of societal powers in the socio technical systems that form the fabric of the Big Data society (Hasselbalch, 2019, p.3). If we accept AI technologies as nodes in our critical infrastructures that manage informational relationships between different societal actors, the distribution of informational resources and the allocation of agency, then we also need to recognize that these technological agents represent our interests and act on our behalf. The core objective of a data interest analysis of AI is consequently to uncover technologically embedded informational asymmetries by questioning their embedded negotiation of interests.

The article begins with an examination of the concept of "data interests". What does this concept mean in the context of current technological change and governance? And which questions do we need to ask to transform these insights into actual ethical design choices? The first premise for the analysis is that the very design of a data system represents power dynamics, or what Langdon Winner (1980) describes as "forms of order" (p. 123) and politics. That developing a technology is also a "structuring activity" that has ethical implications when creating by design "wonderful breakthroughs by some social interests and crushing setbacks by others" (p. 125). An applied ethics (Values by Design/VSD) approach allows us to isolate the very design phase of a data system to analyse the way in which different data interests are negotiated by design. Though a VSD approach is primarily concerned with the identification of transcendental values in design and among stakeholders. A Science of Technology (STS) and infrastructure studies framework on the other hand treats the power (moral or political qualities) of technologies as dynamic concepts in constant negotiation with societal factors. The comprehensive perspectives from these studies therefore form the second premise for the analysis where the design of AI is considered within a general infrastructural macro development that is part of society at large, evolving laws and policies, economy and culture. One of the key strengths of an STS perspective on technological change lies in its recognition of the many modes of governance and actors involved in technological change (from mundane to institutional, from human to artefacts) (Epstein et al, 2016, p. 6). Here, I consider discourse as a mode of governance, and in the context of AI governance, a key factor in the shaping of the conceptual space within which we perceive AI's opportunities and risks. Therefore, following the discussion of data interests, I move on to a qualitative reading of the European Commission's High Level Group on AI's (AI HLEG) Ethics Guidelines for Trustworthy AI published 8th of April 2019[1]. This document does not directly ad-

[1] I was myself a member of this group and took active part in the development of the document. This article is generally informed by a study that I am conducting on data ethics in governance and technology development in the period 2017-2020 based on active participation in various "internet governance" initiatives. I was for example also

dress data interests, however a negotiation of various data interests and attempts to resolve conflicts between them can be discerned in the metaphors concerning data in the text. From this reading, I derive a data interest analytical framework that can be used in the design phase of an AI technology.

Power by Design: What is a Data Interest?

"(...) we needed to talk to the union in order to use their data, but then in the end we didn't use it."

AI developer, 2019

What is AI? [2] The answer is not unequivocal. Because AI is in fact not just one thing. While it is a technological endeavor indeed, AI is also business, politics, economics and culture depending on context and discourse. In the following, I propose that we focus on AI as data to help shape and direct its evolution and deployment in society (Hasselbalch, 2018). In truth, many times, we are distracted by and disempowered by one dominant discourse on the nature of AI. AI as an uncontrollable future scenario (i.e. as autonomous free agents, the inescapable software evolution of humankind, guarded business trade secrets or governments' national trademarks) is a powerful discourse that limits us in what we think we can do with AI and how we can shape its trajectory. A data interest analysis, on the other hand, understands AI technologies as complex data processing systems and data design. Nothing more and nothing less. Yet, data design that forms a type of embodiment of order in our information realities.

As key starting point, I define a data interest 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 analysed by the AI technology. By understanding a "data interest" in this way, I emphasise an often very explicit formulation in public debates about political, business and civil society motives and intentions regarding digital data. Data interests can be many, they may compete, or be aligned, explicitly, implicitly or not considered at all in data design, and can represent micro individual stakeholder objectives, values and needs (from those of the developers to those of the users) or macro societal and cultural requirements or politics (from new legal frameworks to dominating cultural sentiments and aesthetics). Crucial to the conceptualization of a data interest is the recognition that it evolves

part of the Danish government's data ethics expert committee (2018) and I am co-founder of the non-profit organisation DataEthics.eu active in the field. This embedded approach has been essential to my understanding of the underlying dimensions of the topic of this article.

[2] I use the term AI generically to address public discourse on the topic. Thus, I do not refer to the "intelligence" of AI (technologically or philosophically), but consider AI as automated decision-making *data intensive* systems as in the definition of AI in the AI HLEG ethics guidelines (2019), p. 38.

in a dynamic social and cultural environment and therefore must be considered as a set of questions we pose throughout the life time of an AI technology's development and adoption.

This understanding of data interests may be illustrated with an example. In 2018, stories about the AI "babysitter" app Predictim ran its course in the news. Predictim uses an "AI scan" (language-processing algorithms and image-recognition software) to analyse social media posts of babysitter job seekers on sites like Facebook, Instagram and Twitter to produce a personality report with a risk score. Only employers have access to Predictim's report. (Harwell, 23 November 2018, Merchant 12 June 2018). At first glance, the interests in the data of Predictim are easily detectable. There are for example the app developers' interest in bettering the AI functionality enriching it with data from social media. From the users of the app's point of view, they have (as employers) interest in the insights provided by the data analysis of the app and then there are the job seekers' interest in keeping their social media data private or even just their interest in reviewing the report based on their data. Other interests in the data of the app came in to public view when the story about the app was brought in the Washington Post (Harwell, 23 November 2018) describing the app's use of job seekers' social media data to predict risky behavior. Following this, Facebook and Twitter banned the app from their portals to prevent access to their users' personal data claiming this contradicted their interest in the data, that is, their terms of use (of data shared on their portals). But Predictim's CEO, Sal Parsa, said the company was not doing anything wrong "Everyone looks people up on social media, they look people up on Google," Parsa told BBC News, "We're just automating this process" (Lee, 27 November 2018). One macro level interest in data that the company did not consider was a shift in the general societal interest in the social implications of big data taking the form of new legal requirements, citizen concerns and changing business ethics (Hasselbalch & Tranberg, 2016). Although Predictim was just doing "business as usual" with digital data, the cultural sentiment had shifted and data interests were transforming alongside rapid technological change transforming mundane data analytics with new AI capabilities.

What is a data interest? In the Predictim case, the public discussion revealed a set of data interests working on different societal levels from micro developer considerations, to business and user group interests and to macro societal and emerging social and cultural sentiments. Not all AI technologies are subject to the same type of public review as in Predictim's case and rarely as explicitly reflected on during the development of an AI technology's design although they should be. But how do we constructively approach the complex set of data interests present in the design of an AI technology?

The "ethics by design" approach has gained momentum in public discourse in recent years in the conext of AI development and policy (Dignum et al, 2018). The conceptualization of the entrenched values of a computer technology design was originally formulated by Friedmann and partners in the 1990s and has since then been further explored in the Value Sensitive Design (VSD) framework (Friedman, Kahn & Borning, 2006, Friedman & Nissenbaum, 1997, Friedman, 1996, Friedman & Millett, L., 1995, Friedman & Nissenbaum, 1995, 2, Friedman, Kahn & Borning, 2006). In VSD the embedded values of a technology are addressed as ethical dilemmas or moral problems to solve in the very design and practical application of computer technologies. For example, in the context of data interests, one can with VSD focus on "privacy" as a value that can be adversely affected by the specific design of a data intensive technology, but that can also deliberately be designed into that technology (as for example a "privacy-by-design" (Cavoukian, 2009) approach to technology design would entail). Interests are here correlated with values that are held by stakeholders and that can be affected by a technology's design. Thus, the aim is to develop analytical frameworks and methodologies to resolve in the design of a computer technology conflicts between the needs and values held by these stakeholders (Umbrello & De Bellis, 2018, Umbrello, 2019).

While looking to resolve conflicts in the very design of a technology is core to a data interest analysis, the VSD emphasis on values per se is also a limit. The very idea that values can be built in to a technology does presuppose an understanding of values as stable and transcendental qualities taking for granted their moral good or harm. Thus, our attention is diverted from the negotiation of interests in their broader social contexts and as a result the act of evaluation and assessment in social contexts is placed in the background.

In contrast, interests as a classical sociological concept are situated in social contexts characterized by structural power dynamics in society. As such the design of a technology could be said to be primarily shaped by this context and thereby the result of the dominating interests in a given society, that is; the hegemony of those in power. With a sociological approach to interests and "interest- oriented-action" (Spillman & Strand, 2013), the focus shifts from the micro stakeholder environment of negotiation of transcendental moral values to the macro context of general societal power struggles between different stakeholder groups representing different social interests. Interests are here "a major determinant of social action", or analytical categories for understanding societal developments (Spillman & Strand, 2013, p. 86). For example, we may consider big technology industries' data intensive technologies as the expression of a logic of the capitalist society. Subsequently, we could explain the design of a data intensive technology as the result of a capitalist logic and a representation of a big data industry's interest in acquiring the "raw material" of their business models with a direct effect on the interests of its users to preserve their right to privacy. In this perspective, the core societal solution would not lie in technical design as such, but in a counter action in social organization; that is in the social processes leading to technology innovation and design.

For a data interest analysis, an understanding of the interplay between technological innovation and general societal power dynamics is important. In particular the emphasis on finding answers to social challenges posed by technological innovation in a social process that can be translated in to a continuous socially situated reflection on embedded data interests throughout the development and deployment of an

AI technology. At the same time a reductionist view on macro interests in big data also diverts our attention from the micro environment of interests embedded in the very design of a data intensive technology. Speaking with developers of AI technologies, one for example will understand how macro and micro interests are constantly at play in the developmental phase of a data intensive technology. One AI developer[3] for example talked about the different interests in the data she uses in her work from micro developers' interests in acquiring the data to fulfill the specific task of for example a narrow AI product, to macro structures of big tech industry interests in data. Often these interests are intertwined and balanced against basic functionalities, constraints and technical issues. As she says: *"AI is like very data hungry, so when we are building something, then we are like thinking what API is there that we can use for this, so what API Google provides, like Bing or Amazon, whatever, what API is there that I can take the data, and make use of this…"* She then continues to explain how most companies don't have the hardware to train their AI models, so they just buy Amazon, IBM or Google cloud services. But she worries about the data she is sharing with these cloud services *"we just say that this is covered by their privacy policy, but we don't know exactly what are the clauses, I don't know, to be honest",* and when in response to her concerns she is queried further about what she gets out of using these platforms as an AI developer, she answers: *"I have good hardware, so I can run things way faster than I would on my computer, so I get the speed"*. In this way, a design choice with a trade-off between the data interests of users, the cloud service providers and the AI model is made. However, later she recalls a case in which the data interests of a group of people within a specific occupation trumps a business client's data interests (I have changed the occupation here to preserve the confidentiality of the developer and company in question): *"I remember there was one case, which …involved one occupation…so it was like there was* florists *involved, we wanted to use some information about their performance, but then there was the union of* florists*, so we had to negotiate with the union of florists in order to use some data about the* florists*, (…). So they were like basically trying to protect the interests of* florists *because then you can kind of infer something about the performance of a particular* florist *which may, I mean we are not gonna use it, but maybe the company who is our client may be able to say something about the performance of the* florist*…"*. When asked how a union got involved, she discloses another design choice based on the negotiation and trade-off between data interests: *"(...) we needed to talk to the union in order to use their data, but then in the end we didn't use it"*. This conversation with a data designer illustrates the importance of a holistic STS approach to technological development that takes into account the dynamic process in which innovation, the engineering of technologies, and the technologies themselves are simultaneously shaped by society and society shaping (Hughes, 1987, p. 51). The very design process of a technology is intertwined with social, economic and cultural problems and, as Bijker & Law (1997) describes it, its potential working or failure is shaped by a range of diverse factors from politics, economics and aesthetics to available tools and resources, professional skills and prejudices that are all "thrown into the melting pot whenever an artifact is designed or built" (p.3). In this way, the future scenarios of AI adoption in society equally become tangible. As Lapenta (2017) states, the future is "not arbitrary but the product of a complex series of decisions and actors that can potentially give shape to a number of differently

[3] Interviewed in 2019

possible, probably, or desirable future scenarios" (2017, Lapenta, p.154) emphasizing present day conditions for technological development as a reflective choice we make. Callon (1987) describes engineers as sociologists ("engineer- sociologists") involved with social analysis and problem solving from the beginning of a process of technological innovation (p.84). He advocates that sociology could gain valuable insights from studying the type of sociological analysis that takes place within engineering processes rather than only studying them as technical endeavours (p.87). But we could also turn this argument around and propose that engineers need a more reflective assessment of their own work as a form of social engineering practice. As Bijker & Law (1997) puts 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).

Langdon Winner (1980) describes technologies as "ways of building orders in the world" (p. 127), 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, but have embedded politics. They are the locus of the distribution of societal powers shaped by "human motives" (p. 124) and embodies "power and authority" (p. 131). Bruno Latour (1992) equally places an emphasis on technological artefacts as "delegated non-human actors" (p. 157) that enforce human laws (which includes "values, duties, ethics" (p. 157)). As such technological artefacts are "strongly social and highly moral" (p. 152) working by prescription (of laws and orders that are "inscribed or encoded in the machine" (p. 177). A technological artefact such as a seat belt, as he describes it, can for example lock our bodies in positions we do not wish to be in. It is designed to do exactly this, does indeed enforce the laws of car safety, and will certainly let us know with an insistent beeping if we are not ascribing to these laws.

In the context of data intensive technological systems, we may equally see ourselves as locked in specific positions prescribed by the moral laws that are inscribed in the different combinations of data sets that either grant us access or reject access to insights that may be transformed into lost or found opportunities**.** When we perform a search online, we initiate a data analytical process based on a specific data design of a language processing algorithm. This will present us with the insights we are seeking in a prioritized list with links to the information we are looking for and we will certainly go through the first couple of links. Whether we act on the information we access, or not, the design of the list of information can still be argued to have embedded "morals" that are inscribed in the algorithm of the search engine.

Bolukbasi et al (2016) examined how societal gender bias in data from news articles is reproduced and amplified in some of the most popular machine learning methods for language processing used in online search engines. Based on the existing bias of the training data, they found that words are organized in clusters of words such as architect, philosopher, financier and similar titles that are grouped together semantically as "extreme he" words, whilst words such as receptionist, housekeeper and nanny are grouped together as "extreme she" words (p. 2). An employer's online search for a candidate, might very well present to her a list of information embedded with this type of preexisting societal bias. The list represents the delegated moral agency of humans and society, which prescribes on her a very specific prioritization of the information she should look into.

We may infer by this that the work of an "information scientist", or what I for the purpose of this article call a "data designer", entails the delegation of moral values and choices, which therefore also ultimately entail the crafting of "people's identities, aspirations, and dignity" (Bowker & Star, 2000, p. 4). To Bowker & Star (2000) classifications of information are never neutral. As such there is always a "moral" dimension (p. 5.) to the work of a data designer who has an important role as the human who inscribes the programming language of the technology's delegated moral agency (Latour, 1992). However, the translation of this type of technological moral agency into moral action and/or implications is not a straight forward process. This is what Latour (2002) refers to as "technical mediators" where technical design is intertwined and transformed into various possibilities in constant negotiation with our actions, the contexts of the laws, culture and society in which we act, and where morality is not just a normative preexisting framework that we are, as morally aware human beings, are obliged to consider and act on. Morality is the exploration of "various means to ends" deeply intertwined with the technical and vice versa (Latour, 2002). Continuing with our example of the search algorithm, which is designed to classify data on which the system infers patterns and make different types of analysis. In a computer scientist's language, these classifications of data must be "accurate" to provide the best analysis. But we may ask what is an accurate classification? If it is an unbiased one, a concern that is increasingly creeping into the public AI discourse, we should also ask, to whom is it unbiased and based on which criteria? Whose interests does the accuracy serve?

Exactly due to the technical mediators, the moral work of data designers cannot be prescribed with a preexisting ethics by design manual, it is contextual and habitual, and therefore should take the form of a continuous reflection, questions and assessments that will help them unveil and make transparent the data interests in their design and explore their design choices' many potential implications for the agency of different sets and forms of often intertwined interests.

Agency is a key concept in theories of interests. As illustrated, in AI technology development trade-offs are constantly made between different interests in data. As a result, the invested data interests are provided with agency or little or no agency in the final product. Agency theory examines informational relations between "agents" that manages the (social and/or economic) interests of "principals" (Shapiro, 2005). Core to agency theory is that principals delegate decision-making authority to agents that are to represent their interests. To provide some examples: An agent could be a real estate agent, and the principal the buyer and/or seller of a property whose primarily economic interests are negotiated in the information contained and shared in contracts and relationships. But an agent could also be a nation state, and the principals 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, the same does a social media company's untransparent harvesting and processing of the personal data of its users. Both forms of asymmetry have a direct effect on 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. 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 interest. Hence, we may consider AI technologies as "moral agents", not with moral agency as such (Adam, 2008) capable on their own of making moral decisions, but as agents in which humans delegate the enforcement of agency of different interests (Latour, 1992). 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" where all interests are served by having an equal amount of information to make rational decisions (Neumann & Morgenstern, 1953). Of course, technological development is always embedded with societal trade-offs (Bijker & Law, 1997) and consequently ethical choice. Information is never perfect. It has a moral dimension, that is; choices that entail trade-offs between interests are made and should from an ethical point of view be made between different interests in the design of a technology. However, without a thorough understanding of the actual interests at play in a data design we cannot make ethical choices, therefore first step is to identify the interests. In the following, I provide a framework to do exactly that.

The Data Interest Analytical Framework

"…Revolution is defined as a cathartic act meant to reveal the political load of the world: it makes the world*; and its language, all of it, is functionally absorbed in this making."*
(Barthes, 1972, p. 147)

Here, I present a framework for a data interest analysis based on the data metaphors embedded in the ethics guidelines for Trustworthy AI published 8th of April 2019 by the European High Level Expert Group on AI (AI HLEG).[4] The guidelines were developed by a group of 52 experts appointed by the European Commission representing stakeholder interests from industry, civil society and academia.

[4] The AI HLEG ethics guidelines are one set out of many AI ethics guidelines published by various groups and initiatives in the years 2018-2019. Jobin et al have identified 84 documents revealing some thematic convergence between them, but also conceding that there is "substantive divergence in relation to how these principles are interpreted; why they are deemed important; what

Although the very text of the guidelines does not directly address data interests, a qualitative reading of the way in which the data of AI is treated metaphorically reveals a negotiation of various data interests and attempts to resolve conflicts between them. The governance framework for the data of AI is an underlying theme throughout the ethics guidelines. Data design is treated as a locus of power and interests, and accordingly, as I will show in the following analysis, becomes an expression of different societal types of agency - of the technology itself, of legal frameworks, of the users, and of deployers and developers of AI.

Lakoff & Johnson (1980) famously described how our conceptualization of the world is structured metaphorically. Understanding one thing in terms of another is not just an arbitrary exercise, they argue, metaphors give shape to and guide our actions, they highlight or hide and importantly direct our focus (p. 10). Inevitably, metaphors "play a central role in the construction of social and political reality" (p.159). But more importantly, we make sense of the world of things "on the basis of our own motivations, goals, actions, and characteristics" (Lakoff & Johnsson, 1980, p. 34). Thus, language is powerful, it naturalizes our views of the world that enforce power dynamics as Roland Barthes (1972) illustrated in his analyses of the myths of human language in all its communicative forms. We may conclude then that also the language we use to describe AI innovation, law, culture, and governance has a "governing" function. That is, it shapes what we think we can do with the technology and how we can design it, govern and direct it.[5]

In the ethics guidelines, five thematic metaphorical clusters are presented representing different societal interests in the data of AI: Data as resource, Data as power, Data as regulator, Data as vision and Data as risk. Within each metaphorical cluster, I derive key questions that may facilitate an exploration of the way in which an AI technology's data design supports and/or represses the agency of specific data interests. The questions should all lead to one all-encompassing question to any data design of an AI technology: would an alternative data design solve conflicts between data interests in a way that serves an ethical purpose?

Data as resource:

The "data as resource" cluster is core to any data interest analysis and therefore here also described first and most extensively. It concerns the very distribution of data resources among involved interests in data design. In the ethics guidelines data is a resource, metaphorically separated from that which it represents

issue, domain or actors they pertain to; and how they should be implemented." (Jobin et al, 2019, p.1). A distinct point of departure for the AI HLEG is the "European interest" emphasised with a direct reference to the European Commission's vision to among others ensure *"an appropriate ethical and legal framework to strengthen European values"* (p.4) as well as a direct reference to European legal and cultural frameworks.

[5] I have previously used this framework of analysis in non-academic talks and writings "Language, Power and Privacy" https://mediamocracy.org/2014/08/26/language-power-and-privacy-talk-at-the-indie-tech-summit-brighton-july-2014/, "Let's talk about AI" https://www.linkedin.com/pulse/lets-talk-ai-gry-hasselbalch

(a person or an artefact). 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 "data as resource" metaphor is a common metaphor in public discourse on big data. The technology critic Sarah M. Watson refers to the dominant metaphors for personal data in public discourse as "industrial", as if it was a "natural resource" to be handled by "large-scale industrial processes" (Watson, no date). Mayer-Schonberger & Cukier (2013) depicts data as the raw material of a new type of industrial age. But the "data as resource" metaphor is actually used to denote to 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 capacity, physical, psychological and social means. The first type of resource mentioned here is tangible and material, the other social and psychological. But both are, what Lakoff & Johnsson (1980) would refer to as "container" metaphors, entities with boundaries that we can handle and reason about (Lakoff & Johnsson, 1980, p. 25). 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 interests in the raw material of their data based business models; political players have interests in providing rich data infrastructures for AI innovation to strive in their regions; the engineer has an interest in the same 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 e.g. the personal data store, or "data trust", movement is representative of). Obviously treating the psychological and social resources of individuals as material resources in an industrial production line represents a conflict in 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 interests 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 requirement, p.17). Closely connected with the principle (ii) "prevention of harm", the protection of data as a social and psychological resource of human beings, is described as *"…protection of human dignity as well as mental and physical integrity."* (p. 12) which 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 a foundation of what is the objective of the guidelines, the development of "Trustworthy AI".

Questions:

Who or what provides the data resource? Who or what has interest in the data resource? How is the data resource distributed? Whose capacity is enhanced by it? How are conflicts between micro level (of for example engineers, the systems, the technology, the company, the institution, the users of the technology) or macro level data interests (digital gaps, citizens and individuals, global competition, technological change, democratic processes) addressed? In cases where personal data is involved, is the individual human being's social and psychological resourcefulness protected and enhanced?

Data as power:

The second cluster "data as power" is closely associated with the first cluster as the distribution of resources also constitutes a distribution of power. Societies are based on balances of power between social groups, states, companies and citizens. A democratic society 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 distributions of power in society. Surveillance studies scholar David Lyon (2014) envisions an "ethics of Big Data practices" (2014, p. 10) to renegotiate a recent unequal distribution of power embedded in the technological big data infrastructures. In the guidelines, changing power dynamics are addressed in the 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 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 *"…the decision must be identifiable, and the decision-making processes should be explicable."*, p. 13) is the basis of *"accountability"* and *"fairness"* (p. 12)(*"Without such information, a decision as such cannot be duly contested."* p. 13). The data design solutions are associated with the concept of *"human agency"* (p. 16) that supports *"…informed autonomous decisions regarding AI systems."* (p. 16) or *"…informed choices in accordance with their goals."* (p. 16).

<u>Questions:</u>

Who is empowered or disempowered by the data access and processing? Whose power does the data agency support? How are conflicts between different data interests resolved?

Data as regulator:

The third cluster "data as regulator" represents the legal enforcement of the power balancing of data interests. Technology design is a type of "regulator" that either protects or inhibits legal values. "Code is law", as law professor Lawrence Lessig (2006) formulates it quoting Joel Reidenberg's concept "Lex Informatica". Reidenberg links technological design choices and "rule making" with governance in general, understanding the two as inseparable, the first enforcing the other (Reidenberg, p. 555). In an ideal constellation, law and technology design supplement each other. The ethics guidelines propose three core components that are prerequisite for Trustworthy AI: 1. Lawful AI, 2. Ethical AI and 3. Robust AI. While it is emphasised that the guidelines do not replace existing law, and therefore the first component "Lawful AI" is only a reiteration of compliance with European legal frameworks, it can be argued that the guidelines still treat the so called "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)."* 2018, 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 for example 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 implementing 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 legal implementation and the 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 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, 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) explanation. Importantly, different types of societal stakeholder interests in the data design may be in conflict with legal frameworks such as the GDPR or the European Fundamental Rights Framework in which the protection of the interests of individuals is a primary consideration. State actors' interests in control, optimization of institutional processes or in national security; Business interests in tracking and collecting data to enhance their data based business models; Scientific interests improving scientific discoveries with big data analytics. The ethics guidelines mention some of the most challenging legal challenges of AI in a European Fundamental rights framework: mass surveillance, discrimination, the undermining of democratic processes and proposes alternative design options embodied in the concept of Trustworthy AI.

<u>Questions:</u>

Does the data design of the AI system protect or inhibit legal compliance? Are there conflicts between big data interests and the interests of a legal framework such as the GDPR? How are they sought resolved in design? (e.g. such as design of personal data control, privacy by design, data minimization, right to explanation)

Data as eyes:

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 see, and how we see it) in digital information infrastructures. The eye is the agency of vision, meaning we may have eyes without vision, though we may not have vision without eyes. In an ideal constellation, vision is an effortless extension of our eyes. However, in a digital information based environment, the instruments (our digital eyes) we use to see and perceive our environment with 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 it exists in. In the ethics guidelines eyes and vision is embodied in data. 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 technology. 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, p.107). 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 (the data design) of an AI technology does exactly that. To provide an example, we might create an AI technology to analyse the data of people on social welfare benefits. This has a dashboard for the public institution's social workers which provides general data statistics, fraud detection and risk scores on individuals. The dash board 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, optimizing work and also supporting citizens. 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.

<u>Questions:</u>

Who or what does the data design provide eyes and vision? To whom are the data processes transparent? And what does the data design see (the training data) and then perceive (how is it instructed to act on training data)?

Data as risk:

In the fifth last cluster "data as risk", data interests are correlated with the act of assessing 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 economy, on businesses and the environment. The risks are the harms of AI that should be anticipated, prevented and managed (… *"AI systems also pose certain risks and may have a negative impact, including impacts which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself.) Adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk."* p. 2). In particular in the third core component ("Robust AI"), a reference to a 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. The sociologist Ulrich Beck described this societal preoccupation with risk prevention and management in his seminal book *The Risk Society* (1992) as an uncertainty produced in within 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, Beck (2014) later states when describing a development of concern with "world risks", they take the shape of "the anticipation of catastrophe" (p. 80) 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)." (p.81) In the guidelines, the data of AI is 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, 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 in itself also has risky properties e.g. it can be *"malicious"* or *"adversarial"*. 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.

Questions:

Who or what could the data design be a potential risk to? Who or what has an interest in preventing and managing the identified risks? Are there conflicts and/or alignments between identified risks and interests in managing the risks?

Conclusion:

"I always end up exactly where I need to be. In spite of the fact that it is rarely where I intended to go…"
(Dirk Gently, "Dirk Gently's Holistic Detective Agency", BBC/Netflix, 2018)

Over a very short period of time there's been an acceleration of the integration of data intensive technologies such as AI in all areas of our lives. A complex and advanced machinery designed to act on data, predict behavior and trajectories and point lives, economies, politics, society in specific directions. AI technologies constitute a distribution of informational resources in society and therefore have a direct influence on the balance of agency among societal actors. We therefore need a to consider their built-in trade-offs between different interests of their data design.

In this article, I have proposed 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. With point of departure in an action oriented framework that studies the distribution of societal powers in the socio technical systems of the Big Data society (Hasselbalch, 2019, p.3), I have built this analysis on a qualitative reading of the European High Level Expert Group on AI's Trustworthy AI ethics guidelines. In this, I identified five thematic metaphorical clusters representing different societal interests in the data of AI: Data as resource, Data as power, Data as regulator, Data as vision and Data as risk. The "data as resource" cluster is core to any data interest analysis as it concerns the very distribution of data resources among involved interests in data design. The second cluster "Data as Power" is closely associated with the first cluster as the distribution of resources also constitutes a distribution of power. The third cluster "Data as regulator" represents the legal enforcement of the power balancing of data interests in society. 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 see, and how we see it) in digital information infrastructures. In the fifth last cluster "Data as risk", data interests are correlated with the act of assessing risks of data design. Within each cluster I formulated key questions that may support an assessment of the way in which data design supports and/or repress the agency of specific data interests. The main question a data interest should always pose is if an alternative data design would solve the identified conflicts between data interests in a way that serves an ethical purpose.

What do I mean by "ethical purpose"? Here, I have described AI technologies as agents in which humans delegate the enforcement of the agency of different interests. This, I have argued, allows us to recognize data design as central to an identification of the ethics by design proper to our ethical purpose and to make choices accordingly. The idea that we can design technologies with ethical purposes in mind, or what is also referred to as the "ethics by design" approach, calls for an emphasis on the moral qualities of technologies, that is; respectively their embedded "values" (as e.g. in a VSD approach) or their "politics" (as e.g. in a STS approach). I have in this analysis chosen to emphasise the political qualities of technologies with the concept "data interests". I chose this focus for my analysis for two reasons:

1) Every day negotiations are taking place between interests in very tangible digital data resources. Data is what the politics of the big data age is all about and also the power struggle we may see represented and reinforced in the data design of AI (from micro engineering of a data hungry technology to global macro political battles over data resources)

2) The very act of moral evaluation and assessment is crucial to ethical action. The idea that "values" can be identified and built in to a technology presupposes an understanding of values as stable and transcendental categories that takes for granted their moral good or harm. "Interests" on the other hand are negotiated and shaped in social contexts. They are held by individuals, groups, companies, institutions, governments and they take many shapes in legal, cultural, economic, frameworks. Importantly, interests may transform, be discarded, if assessed not to serve an overall ethical purpose situated in a specific social and historical context. An ethical purpose is in this perspective also a choice equally shaped by our conditions.

This data interest analysis' condition was a choice. It took point of departure in the European AI HLEG ethics guidelines grounded in the ethical choices already embedded in a European fundamental rights legal framework regarding the respective balance between the powers of individual citizens and other powers in society. It prioritizes the "human interest" over other societal interests (commercial, state, scientific) in the development and integration of big data technological systems in society (what is also increasingly referred to as the "human centric approach", see Hasselbalch, 2019). This is the perspective from which we should always interpret law, innovation practices, business practices and technology development. Always looking at our practices from the interest of the human being and the future of our intelligent technological data environment as one that can be and should be shaped and designed with this interest at the forefront.

Bibliography

Adam, Alison (2008) "Ethics for Things". Ethics and Information Technology 10, no. 2–3 (September 2008): 149-54.
Doi: https://doi.org/10.1007/s10676-008-9169-3

Barthes, Roland (1972) *Mythologies, Selected and translated from French by Annette Lavers* New York: The Noonday Press Farrar, Straus and Giroux.

Bijker, E. Wiebe, Law, John (1997) *Shaping Technology/Building Society Studies in Sociotechnical Change*. Cambridge, Mass.: 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" , NIPS 2016

Bowker, Geoffrey C., & Star, Susan Leigh (2000) *Sorting Things out: Classification and Its Consequences. Inside Technology*. Cambridge, Mass: MIT Press.

Brighenti, Andrea Mubi. "New Media and Networked Visibilities". In Brighenti, Andrea Mubi *Visibility in Social Theory and Social Research*, 91–108. London: Palgrave Macmillan UK, 2010.
Doi: https://doi.org/10.1057/9780230282056_4

Callon, M. (1987). Society in the making: the study of technology as a tool for sociological analysis. In Wiebe E. Bijker, Thomas P. Hughes, & Trevor Pinch (Eds.), The Social Construction of Technological Systems (pp. 83-103). Cambridge, MA: MIT Press.

Cavoukian, Ann (2009) "Privacy by Design The 7 Foundational Principles", Canada: Information & Privacy Commissioner.
Retrieved from https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf

Deleuze, G. (1992) "Postscript on the societies of control". October, 59, p. 3-7.
Retrieved from http://www.jstor.org/stable/778828

Dignum, Virginia; Baldoni, Matteo; Baroglio, Cristina; Caon, Maurizio; Chatila, Raja; Dennis, Louise; G´enova, Gonzalo; Kließ, Malte; Lopez-Sanchez, Maite; Micalizio, Roberto; Pav´on, Juan, Slavkovik, Marija; Smakman, Matthijs; Steenbergen, Marlies van; Tedeschi, Stefano; Torre, Leon van der; Villata,

Serena; de Wildt, Tristan; Haim, Galit (2018) “”Ethics by Design”: Necessity or Curse?”’ In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES ’18*, 60–66. New Orleans, LA, USA: ACM Press, 2018.
Doi: https://doi.org/10.1145/3278721.3278745.

Edwards, P. (2002) “Infrastructure and modernity: scales of force, time, and social organization in the history of sociotechnical systems”. In Misa, T. J., Brey, P., & A. Feenberg (Eds.), *Modernity and Technology* (pp. 185-225). Cambridge, MA: MIT Press.
Elish, M. C., and danah boyd. ‘Situating Methods in the Magic of Big Data and Artificial Intelligence’. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 20 September 2017.
Retrieved from: https://papers.ssrn.com/abstract=3040201

Epstein, D. & Katzenbach, C. & Musiani, F. (2016). “Doing internet governance: practices, controversies, infrastructures, and institutions”. Internet Policy Review, 5(3).
DOI: 10.14763/2016.3.435

Friedman, B., Kahn, P. H., Jr., and Borning, A. (2006) “Value Sensitive Design and information systems”. In P. Zhang and D. Galletta (eds.), *Human-computer interaction in management information systems: Foundations*, 348-372. Armonk, New York; London, England: M.E. Sharpe.
Retrieved from: https://vsdesign.org/publications/pdf/non-scan-vsd-and-information-systems.pdf

Friedman, B. and Nissenbaum, H. (1997) “Software agents and user autonomy. Proceedings of First International Conference on Autonomous Agents”, 466-469. New York: ACM Press. 1996
Retrieved from https://vsdesign.org/publications/pdf/friedman97softwareagents.pdf

Friedman, B. (1996) “Value-sensitive design”. ACM interactions, 3(6), 17-23.
Retrieved from https://vsdesign.org/publications/pdf/friedman96valuesensitivedesign.pdf

Friedman, B., and Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems, 14(3), 330-347. 1995
Retrieved from https://vsdesign.org/publications/pdf/64_friedman.pdf

Friedman, B., and Millett, L. (1995, 1) “"It's the computer's fault" -- Reasoning about computers as moral agents”. Conference Companion of CHI 1995 - Conference on Human Factors in Computing Systems, 226-227. New York: ACM Press.
Retrieved from: https://vsdesign.org/publications/pdf/friedman95fault.pdf

Friedman, B., and Nissenbaum, H. (1995, 2) “Minimizing bias in computer systems”. Conference Companion of CHI 1995 Conference on Human Factors in Computing Systems, 444. New York: ACM Press.
Retrieved from: https://vsdesign.org/publications/pdf/friedman95minimizebias.pdf

Hasselbalch, G., & Tranberg, P. (2016). *Data Ethics. The New Competitive Advantage*. Copenhagen: Publishare.
Retrieved from: https://dataethics.eu/wp-content/uploads/DataEthics-UK-original.pdf

Hasselbalch, G. (2019). “Making sense of data ethics. The powers behind the data ethics debate in European policymaking”. Internet Policy Review, 8(2).
DOI: 10.14763/2019.2.1401

Hasselbalch G. (2018) "AI: The Data Ethics Perspective", (translation) in Jørgensen, Rikke Frank, Kofoed, Birgitte, Eksponeret, Copenhagen: Gads Forlag.
Retrieved from: https://dataethics.eu/artificial-intelligence-opens-a-myriad-of-data-ethics-questions/

Harwell, Drew (November 23, 2018) “Wanted: The ‘perfect babysitter.’ Must pass AI scan for respect and attitude”, Washington Post.
Retrieved from https://www.washingtonpost.com/technology/2018/11/16/wanted-perfect-babysitter-must-pass-ai-scan-respect-attitude/

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). Cambridge, MA: MIT Press.

Lakoff, George, Johnson, Mark (1981) *Metaphors We Live By*. Chicago: The University of Chicago Press.

Latour, Bruno (1992) “Where are the Missing Masses? The Sociology of a Few Mundane Artifacts”, in Bijker ,Wiebe E. and Law, John (eds.), *Shaping Technology/Building Society: Studies in Sociotechnical Change* pp. 225–258. Cambridge, Mass.: MIT Press, 1992.
Retrieved from: http://www.bruno-latour.fr/sites/default/files/50-MISSING-MASSES-GB.pdf

Latour, Bruno, and Couze Venn (2002). “Morality and Technology”. Theory, Culture & Society 19, no. 5–6 (1 December 2002): 247–60.
Doi: https://doi.org/10.1177/026327602761899246.

Lapenta, Francesco (2017) "Using technology-oriented scenario analysis for innovation research" in *Research Methods in Service Innovation*. Cheltenham: Edgar Allen Publishing Limited.

Lee, Dave (27 November 2018) "Predictim babysitter app: Facebook and Twitter take action", BBC News.
Retrieved from: https://www.bbc.com/news/technology-46354276

Lessig, Lawrence (2006) *Code version 2.0*. Cambridge: Basic Books

Lyon, D. (2010). "Liquid surveillance: the contribution of Zygmunt Bauman to surveillance studies". International Political Sociology, 4(4). (pp. 325-338).
Doi:10.1111/j.1749-5687.2010.00109.x

Mayer-Schonberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Neumann, John Von, Morgenstern, Oskar. *Theory of Games and Economic Behaviour*. Princeton: Princeton University Press.

Pasquale, F. (2015). *The Black Box Society – The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press

Shapiro, Susan P. (2005) "Agency Theory" Annual Review of Sociology 31, no. 1 (August 2005): 263–84.
Doi: https://doi.org/10.1146/annurev.soc.31.041304.122159.

Spillman, Lyn, and Michael Strand (2013) "Interest-Oriented Action". Annual Review of Sociology 39, no. 1 (30 July 2013): 85–104.
Doi: https://doi.org/10.1146/annurev-soc-081309-150019.

Reidenberg, Joel R. (1997) "Lex Informatica: The Formulation of Information Policy Rules through Technology". Texas Law Review, 43.

Watson, Sarah (no date): "Data is the New "…"", Dis Magazine.
Retrieved from http://dismagazine.com/blog/73298/sara-m-watson-metaphors-of-big-data/

Winner, Langdon (1980). "Do Artifacts Have Politics?" Daedalus, 109(1), 121-136.