### Abstract

Social media platforms increasingly use powerful artificial intelligence (AI) that are fed by the vast flows of digital content that may be used to analyze user behavior, mental state, and physical context. New forms of AIgenerated content and AIdriven virtual agents present new forms of risks in social media use, the harm of which will be difficult to predict. Delivering trustworthy social media will therefore be increasingly predicated on effectively governing the trustworthiness of its AI components. In this article, we examine different approaches to the governance AI and the Big Data processing that drives it being explored. We identify a potential overreliance on individual rights at the expense of consideration of collective rights. In response, we propose a collective approach to AI data governance grounded in a legal proposal for universal, nonexclusive data ownership right. We use the Institutional Analysis and Development (IAD) framework to explore the relative costs and benefits on stakeholders in two use cases, one focused on digital content consumers the other focused on digital content knowledge workers. Following an analysis that looks at selfregulation and industrystate coregulation, we propose governance through shared data ownership. In this way, future social media platforms may be able to maintain trust in their use of AI by committing to no datafication without representation.

### Keywords

trustworthy AI, data ownership, social media, collective bargaining

# Introduction

Social media platforms increasingly use powerful artificial intelligence (AI) that are fed by the vast flows of speech, video, and sensed data that platforms capture to analyze user behavior, mental state, and physical context. New forms of AIgenerated content and AIdriven virtual agents that are rapidly being integrated into social media platforms. These may be accompanied by new forms of risks in social media use, the harm of which will be difficult to predict, for example, from manipulation of audio and video content such as deepfakes, highly persuasive bargaining/sales agents, accurate lie detection or contextsensitive microtargeting of content. Delivering trustworthy social media will therefore be increasingly predicated on effectively governing the trustworthiness of its AI components.

Some prospective requirements for governing trustworthy AI, such as nondiscrimination, respect for privacy, robusthuman autonomy, and transparency may conflict with the business motives driving AI use in social media. These will present major objectives in designing effective AI governance regimes and may require new innovative forms of governance. However, many strong trends in the development of AI may impede the successful implementation of new governance approaches. Confounding trends include the accelerating pace of AI innovation, especially for organization with access to vast data stores; the diminishing barriers to entry in terms of AI skills and computing power; the transnational nature of AIdriven service provision; the opaque nature of modern machine learning algorithms; and the immense technical and legal expertise that the large digital platforms providing deploying AIdriven social media can bring to bear on any regulatory conflicts. Regulatory approaches to meeting trustworthy AI requirements may also be impeded politically and socially by its perceived chilling effect on AI innovation, especially as AI capabilities to be seen as a strategic national economic and security asset. AI governance regimes that require centralized state regulation and corresponding AI regulatory expertise will therefore require strong political drive to establish and to scale at a pace that can keep up with AIs accelerating impact on society.

In this article, we examine different approaches to the governance AI and the Big Data processing that drives it being explored. We identify a potential overreliance on individual rights at the expense of consideration of collective rights. In response, we propose a collective approach to AI data governance grounded in a legal proposal for universal, nonexclusive data ownership right. We use the Institutional Analysis and Development (IAD) framework to explore the relative costs and benefit to actors in two use cases, one focused on digital content consumers the other focused on digital content knowledge workers, and propose further routes for analysis.

# Governance for Trustworthy AI

AI trustworthiness has risen to prominence in public policy discussion from 2017, popularized by the widespread signup of prominent AI practitioners to the Principles developed by the Asilomar conference on Beneficial AI (Future of Life Institute, 2017). Policy work has built to an extent from existing work on accountable algorithms, which was concerned with fairness and bias in AI applications and the role of transparency and accountability to counteract the opaque nature of modern machine learning algorithms such as deep neural networks (Kroll et al., 2016; Mittelstadt et al., 2016).

In the context of social media, an important work is that of Latzer et al. (2016), which analyses the economics and governance options for what they refer to as algorithmic selection technologies, that is, those that mediate and form the digital reality experienced by individuals and by society. This class of technologies have become dominated by AI techniques and underpin the core functions of social media applications, namely search, aggregation, surveillance, predictions, filtering, recommenders, scorers, content generation, and resource allocation. Risks of algorithmic selection are identified as:

* Manipulation of individuals or groups
* Diminishing variety that creates biased views and distortion of reality
* Constraints on communication and freedom of

expression

* Threats to privacy and data protection rights
* Social discrimination
* Violation of intellectual property rights
* Impact on the human brain and cognitive capacity
* Algorithmic power over human behavior and development

They outline the prospects for different forms of governance of AI that may be followed to reinforce benefits and mitigate the perceived ethical and societal risks of employing AI technology, namely marketbased, selforganization, selfregulation, and coregulation. Marketbased approaches rely on suppliers and customers, especially in consumer markets, applying pressure on organizations to respect their value systems, or risk losing their business. However, this relies on consumers and market partners having a clear understanding of potentially harmful business behavior, whereas may players in the data value chains feeding AI applications operate in obscure data brokerage roles that do not interface directly with consumers. Also, as large digital platforms benefiting from network effects often occupy near monopoly positions there is a high barrier for exit for consumers.

Selforganization involves an organization setting itself principles or standards to follow in developing and deploying AI solutions. These may be administered through internal ethics boards that must be staff with appropriately qualified personnel and sufficiently empowered by senior management to make recommendations that may come into conflict with existing business objectives. Several large digital platform companies such as Microsoft and Google already have such boards in place. However, they suffer from limits to the transparency in their decisionmaking, as they may need to deal with commercially sensitive information or plans.

Selfregulation involves the development of codes of conduct, industry standards, quality seals and certification bodies, ombudsman schemes, and ethic committees across an industry sector. Selfregulation works better in mature sectors characterized by players with a shared outlook on maintaining public and market trust. The Institute of Electrical and Electronic Engineers (IEEE) has commenced as number of initiatives aimed at improving professional practice and trustworthiness in AI technology (Adamson et al., 2019). Significant among these IEEE initiatives has been the Global Initiative on Ethics of Autonomous and Intelligent Systems, which has conducted a wideranging iterative expert consultation on Ethically Aligned Design (EAD), the First Edition of which was launched in February 2019 (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). This sets out general principles for ethically aligned design, highlights a range of issues, and makes recommendations for organization governance, professional practice, government policy, and further research. Many of these issues are now being addressed by the P7000 series of IEEE standards, to provide specific standards for organizations, ranging from process models for ethical design, guidelines on transparency, privacy, bias, and wellbeing, and specific domains for ethical AI including child data, nudging, trustworthy news, and facial analysis. Industry ethical standards rarely involve independent enforcement mechanisms and have been criticized as an anticompetitive means for larger incumbents to raise barriers to new market entrants (Calo, 2017).

These limitations of selfregulation therefore give rise to proposal for coregulation, where industry behavior is impacted by statutory regulation and other government actions such as tax incentives, state funding for research and mitigation measures, and government procurement policy. Such state action has to balance the realization of the potential economic and societal benefits of AI with the anticipated risks. Several governments have started to produce policy documents about the tradeoffs necessary between realizing the benefits of AI. A 2018 review of European national and transnational regulatory proposals identified policy activity (Access Now, 2018). It conducted a gap analysis against criteria related to transparency, accountability, privacy, freedom of conscience and expression, equality and nondiscrimination, due process in law, data protection and user control, collective rights for free press and election, economic impact and the future of work and AI in weapons. It identifies the European Unions (EU) move to use an ethical grounding to differentiate Europes approach to AI, contrasting with the laissezfaire approach of the United States and centralized governmentdriven approach of China, building on the existing commitment to Europeanwide regulation already implemented via the General Data Protection Regulation (GDPR) (EU, 2016), but identifies several gaps AI policy across Europe. Of relevance to AI in social media, gaps are identified in terms of the disparity between standards to which public and private organizations are held, especially given the dominant role of the private sector in developing AI technology. This publicprivate disparity also identified as an issue in relation to big data and AI ethics in a US study (Metcalf et al., 2016). The comparative analysis of EU policy work also highlights relevant lack of policy outside of Germany related to freedom of conscience and expression, both areas heavily impacted by search bubbles and personalized selection of social media news feeds. The profiling of individual is also a highly relevant area, given the inherent power of AI in pattern matching and the vast amounts of personal behavioral data accessible via social media. Though European data protection authorities can interpret such inferred data as personal data, potential for bias exists outside the remit of GDPR when AI application profile places or other nonpersonal characteristics rather than people directly. Gaps are also identified in the purpose limitations of data protection in the face of AIs power in identifying previously unanticipated patterns, such as those inherent in large social media data set. Furthermore, the collective rights to free press and elections are not address in any of the analyzed national policies, despite these being area now being heavily impacted by social media. Attention to the impact of AI on the future of work is identified across the analyzed policies, but primarily in the form of pragmatic of guidance on retraining, and without reference to labor rights. This analysis of European AI policy proposals identifies some important general issues in relation to AI ethics:

* + A general wait and see attitude is prevalent, driven in part by a desire to not to stifle AI innovation and investment. It is also in part driven by uncertainty arising from the rapid progress of AI and therefore the potential for technical solution to emerge (i.e., relying on marketdriven measure), which may quickly render legislation redundant.
  + A lack of consideration of collective issues in identifying and mitigating the risks of AI, rather than relying solely on individual rights and data protection frameworks.
  + The relationship between ethics and legislation, where ethical guidelines may prove useful in reducing harm while appropriate regulation is developed or where the adoption of guidelines simply avoids addressing the difficult issues of regulation, or as pointed out in Calo (2017), can be used by large entities to project a solution without submitting to external enforcement.

These general trends seem to persist in the most recent detailed governmental work on trustworthy AI. The EU has recently established a set of ethical guidelines for achieving trustworthy AI that aims to foster research, reflection, and discussion, but which is positioned as a complement to possible future EU regulation of AI technology (European Commission, 2019). This presents ethical AI principles derived from the EU Charter of Fundamental rights, but consistent with the Access Now review, it focused on the fundamental rights assigned to individuals in relation to chapters on dignity, freedom, and justice but says little in relation to collective rights under the chapter of solidarity including the rights of workers for information, communication, and collective bargaining (Official Journal of the European Communities, 2018). Similarly, recent recommendation on AI from the Organisation for Economic Cooperation and Development (OECD) lays out a similar set of principles that should be followed in the development and use of AI systems (OECD/LEGAL/0449, 2019). However, though it recommends that government should review and adapt policies related to innovation and competition of trustworthy AI, it does not make any concrete proposals on the form of regulation. There have been some proposals for possible coregulatory structures such as new national regulatory bodies responsible for the governance of big data and its use for algorithms (Tutt, 2016) as well as internal ethics boards that may help organizations implement best practice (Calo, 2013; Polonetsky et al., 2015). However, there seem to be few concrete plans in place to establish the public regulatory bodies and standards needed to implement coregulation. This is in part due to uncertainty about the direction and impact of AI technology, but also in part due to the lack of relevant technical expertise available to public policy bodies in comparison with the large digital platforms and other major private adopters of AI (Calo, 2017).

In summary, while the risks of applying AI to social media data may be profound, marketbased, selforganization, selfregulation, and coregulation solutions to AI governance all face challenges in balancing the rapid development of complex AI technology in service of maximizing shareholder value and realization of the societal benefit of AI, with the need to track and mitigate the societal risks the stream of AI innovation present. In the next section, we therefore examine an alternative strand of AI governance proposals that takes a more collective approach by those involved in the provision of data that drives AI, rather than requiring external inspection of algorithmic behavior.

# Collective Governance Approaches for AI Trustworthiness

Another, earlier, thread of research that has had a visible influence on analysis of trustworthy AI is work on the ethics of Big Data, especially in relation to privacy and data protection concerns (Metcalf et al., 2016; Mittelstadt & Floridi, 2016; Richards & King, 2014; Zwitter, 2014). A common theme here is the criticism of the notice and consent mode of personal data selfmanagement that is prevalent as a requirement of most data protection regulation. Privacy researchers already understood that individuals were poorly equipped to fully understand and anticipate the impact of personal data usage to which they asked to consent (Solove & Washington, 2012). This had led to the assessment that the predominant current model of making use of personal data based on initial consent may not be ethically justified and therefore requires some *collective* rather than individual engagement with organizations doing so (Sax, 2016). Big Data analytics, which was later enhanced by rapidly improving AI capabilities in pattern matching, increasingly delivers societal level impacts, rendering individual consent an even less effective mechanism for governing negative consequences of data processing. Different approaches have been proposed to rebalance of this asymmetry of information about the implications of personal data processing. In Rahwan (2018), the case is made for AI governance with societal level engagement through a societyintheloop model. This consists of a number of mechanisms including crowdsourcing in the articulation of values related to AI, introduction of professional algorithm auditors and a new class of algorithm monitoring software, but acknowledges the need for new public institution needed to make this a reality. An alternative suggestion is to treat personal data as a valuable resource shared between the individuals providing it and the organizations collecting and using the data (Tene & Polonetsky, 2013). However, while in the EU such a feature has been included as the right to portability of GDPR (De Hert et al., 2017), it is unclear whether this will result in any compensating realizable value accruing to an individual at a level that would outweigh the transaction cost the individual incurs in managing the additional data sharing. More significantly without an enforceable legal mechanism for realizing the value of data, any transactions would be performed on terms dictated by the company capturing the data in the first place.

# Governance Based on Shared Data Ownership

The problems constraining the leverage of data supply in the governance of AI is a symptom of a broader problem identified in EU policy deliberations on facilitating data sharing to enhance the datadriven economy (European Commission, 2018) related to the legal transaction costs of sharing data. This derives from the complex and multifaceted legal nature of data as identified in report on data ownership rights in the EU (Van Asbroeck et al., 2017a). Data ownership rights to the extent they exist are therefore grounded in a complex range of different legislation, and it is this complexity that impedes the development of data sharing agreements, especially when the respective current and future value of data to the sharing parties is not well understood. There are some limited ownership rights to data through copyright when a level of creativity or originality can be demonstrated, and in Europe through the Database Directive is sufficient effort in the structuring of data can be demonstrated to claim sui generis rights. GDPR offer some specific rights, but only for personal data, and not in the form of familiar ownership rights that can be traded or bequeathed to descendants (the data portability right offers some small move in this direction). Companies may also exert ownershiplike rights over information in relation to patented inventions and through trade secrets.

One proposal resulting from this analysis proposes a new EU data ownership right (Van Asbroeck et al., 2017b). This proposal suggests the nonexclusive ownership on any data for any party that can show a traceability log of having produced or contributed to the data. This is combined with an obligation on such rightholders to share data on fair, reasonable and nondiscriminatory (FRAND) terms provided that this does not conflict with the normal exploitation of the data or if it unreasonably prejudices the legitimate interests of a rightholder. The nonexclusive nature of the ownership right in this proposal reflects the nonrivalrous, nonexclusive and inexhaustible nature of data and the objective to facilitate, rather than restricts its sharing and reuse. The traceability requirement provides a simple but clear opt in mechanism that can be applied to data at different scales, from an individual piece of data, such as a like on a social media post or a sentence translation used to train a machine translation facility, up to entire data sets. Technical standards are already available for capturing such trace logs in an interoperable format, while new distributed ledger technologies offer secure mean for sharing immutable trace logs, greatly simplifying the implementation of the traceability requirement.

Another critical analysis of the data property in the EU highlights the need for transparency in identifying to whom data property rights belong and how this may be confounded by the claims of more than one person, for example, in cases of DNA, group photos, or social media consumer behavior profiles (Purtova, 2017). This means that suggestions related to collective management of data rights, such as the notion of collective consent outlined in the study by Bygrave and Schartum (2009), may struggle to define the boundaries of the group being represented at any point in time. Separate to considerations of data ownership, Mantelero (2016) not only identifies the dynamic boundary of groups in considering data protection concerns of AI, but also the lack of awareness that groups members would have of each other or of shared risks of group discrimination, for example, of lowincome worker driving unsocial hour being grouped with young people driving for socializing. This motivates a proposal for data protection regulators, rather than representative of specific group to play a role in conducting mandatory exante multistakeholder risk assessment of any new group profiling activities. This however seems to be beyond the current remit of regulators, at least under GDPR in Europe, and would have to overcome governments reluctance and limited technical ability to intervene systematically in commercial AI innovation activities.

Data are often referred to as the fuel on which the effectiveness of AI is based, and the impact of any AI application grows in step with the relevance and volume of data available to train the AI and to be processed by it. Our contention is therefore that the stakeholders involved in the sourcing of data for an AI system are often the most representative of those impacted by that AI system. Further if data sources act collectively, they may use their key role in data supply as strong leverage in negotiates that balance between commerciallydriven AI data processing design and the reduction of risks both to the individuals providing data and to other connected stakeholders with whom they empathize, for example, family, friends, groups with specific, sometime minority, characteristics. By grounding representation broadly on data ownership, we hope representation will extend beyond those subject to the specific regulation of data protection, but other groups including knowledge workers, who also have a stake in monitoring and mitigating harms from data gathering and processing by AI.

Therefore, we propose that the obligation to share data under FRAND terms in the data ownership proposal from (Van Asbroeck et al., 2017a) be made conditional in data sharing contracts on safeguarding measures to deliver trustworthy AI. Such conditionality provides an opportunity for new forms of model contracts to be developed that can demand the implementation of trustworthy AI requirements prior to data be shared for the training of or processing by the AI system in question. This approach would enshrine trustworthy AI requirements into the legal contracts that provide access to the data, rather than relying on external regulation of the AI system during its development and operation. Therefore, the agreement on implementing trustworthy AI requirements works in concert with the flow of data that drives AI innovation and commercialization, rather than acting a separate regulatory system that may serve to impede innovation. Implementation of such a data ownership right would therefore introduce shared data ownership as a alternative means for governing trustworthy AI as a complement to marketbased, selforganization, selfregulation and coregulation approaches reviewed above.

However, as discussed above in relation to data protection, individuals are at severe disadvantages in agreeing terms and conditions for sharing data with organizations due to the asymmetry in the technical understanding of data processing and the asymmetry of power due to the relative value of the individuals data compared with that offered by an organization processing data at scale. This would also therefore put individuals exercising data ownership rights at a disadvantage when agreeing trustworthy AI safeguarding terms in a datasharing agreement. While the simple nature of the new proposed data owner right coupled with the practices developed in legislating for intelligible personal data sharing contracts under GDPR may help alleviate the asymmetry of understanding, it does not address the asymmetry in the value being exchanged, that is, individuals would have very little leverage in negotiating strong trustworthy requirements with organizations already providing value from large amounts of information from other sources. However, data rightholders may be able to redress this power asymmetry by *acting collectively*.

Analyzing mechanisms for encouraging and supporting collective action in negotiating shared data ownership therefore requires a shift in how we consider data sharing. A move from considering the sharing of data as a transfer of value to considering it as the management of a shared resource may provide a better grounding for considering the role of all the relevant stakeholders in the governance of data flows and, consequently, of the AI systems that depend on these flows.

By considering the process of developing and deploying AI systems as part of a shared data resource management activity, we are able to compare different approaches to governing this activity through the Institutional Analysis and Development (IAD) framework (Ostrom, 2011). The IAD framework is a tool that assists in designing the institutional structures for the sustainable governance of complex shared resource. The IAD can help build a shared understanding of resource dynamics, the diverse interests within the community of users and producers, and the costs and benefits of different governance structures. It can therefore also be used in comparing different institutional design for a shared resource governance setting. While originally developed for analyzing the governance of share physical resources such as fisheries, it has also been found application in the analysis of shared digital resources, referred to as knowledge commons (Hess & Ostrom, 2007). Such institutional analyses have been used to analyze new forms of digital resource collection, aggregation, and processing, for example, image and geospatial data (Alvarez Leon, 2016). A commonresource pool approach inspired by Ostroms theories is identified as a way to focus debate on the governance of data processing for AI onto the social and political dilemmas it presents (Taylor & Purtova, 2019).

We use the IAD to compare patterns of interaction and likely outcomes from potential AI governance approaches based on selfregulation, coregulation, and shared data ownership. We do this for two action settings related to two classes of AI dataproducing stakeholders: (1) digital content consumers whose behavior is profiled by AI to target further content to them, and (2) knowledge workers whose processing of data in a knowledgerich task contributes to the development for AI to automate of that task. An example of the first action setting could be users of a video streaming platform such as YouTube, where profiles of viewing behavior guide the recommendation of subsequent video content to users. An example of the second action setting could be online customer service agents whose text chat interaction with customers together with customer service system interaction is logged to provide data for automated customer service chatbot. However, in this article, we will focus on the example of translators engaged in postediting of machine translation output to improve its qualify, the result of which is recorded as parallel text used to further train machine translation AI, for example, in the form of statistical or neutral machine translation.