POSITION PAPER

# Digital Ethics: Data, Algorithms, Interactions

Zeynep Engin

Algorithms and digital systems are increasingly taking the role of 'artificial persons' – hence becoming both the subjects and objects of regulation and policing. This summary paper provides an overview of the emerging 'digital ethics' field from a system design and engineering perspective. The objective is to lay out the critical questions and the current research directions that are likely to shape a new 'ethics engineering' profession, which will have significant impact across all sectors.

## 1. 'Digital Ethics' as an Engineering Problem

The future is already here – it is just not fair, safe, legal and transparent enough[1]. Ubiquitous data collection by large multinational companies and governments; increasing deployment of automated decision-making systems operating on massive amounts of both historical and real-time data, producing instant predictions and life-altering decisions; and 'intelligent' assistants, robots and devices interacting with each other and with humans – have all brought us into new *techno-social* spaces and *cyber-physical* realities. As a result, already complex ethical questions, historically debated within philosophy, legal and social science domains, have also become core subjects of interest within the computer science and engineering communities.

*Table 1: Social & Political Concerns around Digital Technologies*

| Aspect | Description | Questions & Examples |
|---|---|---|
| **Democratic** | • Legitimacy and transparency<br>• Asymmetry of information<br>• Public manipulation<br>• Freedom of speech vs censorship<br>• Alternative concepts for democracy | • Algorithmic decision-making<br>• Big Brother state? Multinational companies?<br>• Fake news, voter behaviour, internet trolls etc.<br>• State control of online content<br>• Liquid democracy? E-voting? |
| **Citizen Rights** | • Ownership & control of data<br>• Deriving benefit from data<br>• Fairness vs personalisation<br>• Accessibility & usability | • Distributed vs privileged access to data & information<br>• State, multinational companies, individual citizens<br>• Service delivery, criminal system, iterated bias etc.<br>• Digital literacy and affordability, citizen interactions |
| **Economic** | • Digital monopoly<br>• Crypto-currencies<br>• Sharing economy<br>• Taxation of digital services<br>• Regulation vs productivity | • Multinational companies & centralisation of the markets<br>• Economic value<br>• 'Uberisation' & peer-to-peer service models<br>• Service type, location, employment etc.<br>• Restricting innovation potential? |
| **Psychology & Perceptions** | • Privacy<br>• Transparency<br>• Skewed realities<br>• Addictive behaviour | • Personal data & digital footprints<br>• Black-box computation<br>• Personalised content & recommendations (TripAdvisor, 2017)<br>• Social media, online games |
| **Security** | • Cyber attacks<br>• Identity fraud<br>• Data stewardship & security breaches<br>• Encryption | • Potential physical damage through IoT systems<br>• Equifax (2017), Ethereum hack (2017)<br>• Facebook/Cambridge Analytica (2018)<br>• Preserving privacy vs safe heaven for crime? |
| **Law & Regulation** | • Legal status of algorithms<br>• Regulation of algorithms<br>• Accountability in algorithmic-decision making | • Algorithms as "artificial persons" ?<br>• Regulation of automated transactions (e.g. smart contracts)<br>• Self-driving car assessing scenarios in unavoidable crash? |
| **Evolutionary** | • Cognitive shifts<br>• Technological singularity | • Loss of orientation (e.g. extensive use of GIS), declining attention span, AR & VR<br>• What if the machines take over? |
| **Environmental** | • Natural resources & sustainability | • Electricity consumption & carbon footprint (e.g. BitCoin) |
| **Social** | • Job losses<br>• Retrofitting into existing infrastructure | • Robots replacing humans?<br>• Legacy systems & re-training the existing workforce |

[1] Adapted from the adage *"The future is already here – it's just not very evenly distributed"* by William Gibson, 1993.

Table 1 summarises a range of social and political concerns around new digital technologies. As illustrated, the ethical questions can be raised from a number of different perspectives. These range from fundamental human rights and democratic processes, to more practical matters such as environmental, or, security issues, and to human evolutionary development (particularly visible around the 'technological singularity' debates). There are also strong tensions at the principal level of scholarly debate, such as the one between, on one hand, the *data minimization principle* (organisations should collect only the data they need to answer specific questions) and, on the other hand the fundamental promise of data mining and big data (that valuable and unexpected insight might be made in large data sets, but only if all data is stored)[2]. Clearly, generating an exhaustive list under the 'digital ethics' topic is a very difficult task especially given the fluidity of the domain, and the dependency on varying factors such as time, demography and viewpoint.

With the European Union's introduction of the General Data Protection Regulation (GDPR), an increase in public conversation around 'data ethics' over the past few years has been observed although it is not yet clear how the key concepts – such as *data ownership*, *right to access*, *right to rectification*, *right to erasure*, *right to data portability* – will be realized in engineering and design terms. This is especially the case given the ever-growing digital ecosystem of connected devices and automated decision-making systems together with the conflicting interests of individuals, public and private sector organisations. The need for a similar legal framework for 'algorithm ethics' is becoming increasingly obvious considering, for example, algorithms for CV sifting, loan applications, and recommender systems. Perhaps less obvious is the need for a framework on 'interaction ethics' to regulate both human-algorithm and algorithm-algorithm interactions – for example Chatbots, intelligent assistants, IoT devices, social family robots, etc. Such ethical concerns must be followed with an exploration of the technology landscape for appropriate responses, clearly also considering their limitations and challenges.

Given the significant potential of computing technologies to help us create 'fairer' and 'more ethical' societies versus their observed unethical and illegal decisions in many of the current use cases, a clear fundamental question that should be asked is the extent to which 'ethics' can be mathematically formulated, programmed and embedded in such technologies. This would essentially help identify the boundaries of the 'ethical-by-design' concept often used in framing legislative and policy discussions; as well as the 'ethics engineering' concept introduced in this paper to refer to engineering for built-in ethical functionality (in situ) of the technology-based products. For a systematic description of the landscape, the following definitions are employed as the basis for further discussions in this paper:

> **Data Ethics –** Ensuring moral and legal conduct in the exploitation and utilisation of data and information related to individuals, systems and assets for both private and public purposes:e.g. digital footprints, government records.
>
> **Algorithm Ethics –** Ensuring moral and legal conduct for algorithmic behavior as algorithms increasingly take the role of 'artificial persons' with potential unethical and illegal decisions and actions affecting humans:e.g. CV sifting, customer decisions.
>
> **Interaction Ethics –** Ensuring moral and legal conduct and trust in human-to-algorithm and algorithm-to-algorithm interactions:e.g. 'intelligent' assistants, IoT systems.

The next section outlines the new digital technologies and some of the contested issues associated with each of them. Then the three core concepts – 'Data Ethics', 'Algorithm Ethics' and 'Interaction Ethics' – are introduced in more detail followed by a short discussion around the legal considerations of digital systems and conclusions.

[2] Thanks to David Hand for raising this point.

## 2. The Technology Landscape

Arguably the current digital transformation is driven by the key technologies of Big Data Analytics, Artificial Intelligence (AI), Internet of Things (IoT), Blockchain, and Predictive and Behavioural Analytics (see Table 2 – also discussed further in [1]). Information Security (InfoSec) should be added to the list as the first step towards safe and secure system design processes. Examining the critical issues associated with each of these technologies would provide an explanatory framework to understand their offerings and limitations, as well as identifying their potential complementary uses for the 'ethical design' of future computing systems.

*Table 2: Data Science Technologies & Contested Issues*

| | |
|---|---|
| **Artificial Intelligence (AI)**<br>**Definition(s):** Ability for computers to learn and make decisions without explicit programming. Knowledge-based systems (KBS): systems that reason – knowledge represented as ontologies or rules rather than via code. Machine Learning (ML): AI program with the ability to learn from data without explicit programming. Other concepts - NLP, Sentiment Analysis, etc.<br>**Contested issues:**<br>• KBS designer assumptions and discrete conclusions<br>• ML dependency on training data<br>• ML black-box computation & no insight into the process | **Big Data Analytics**<br>**Definition(s):** Analysing large and varied datasets to uncover hidden patterns, unknown correlations, customer preferences, etc.<br>**Contested issues:**<br>• Data access, quality, representativeness, unstructuredness<br>• No theory: correlation not causation<br>• Potential privacy breaches through the use proxies and linking multiple datasets<br>• Streaming data storage and processing<br>• Real-time data availability & risks<br>• In-memory analytics: processing at server's RAM? |
| **Predictive and Behavioral Analytics**<br>**Definition(s):** Providing insights into the actions of people to help make informed decisions. Behavioural Analytics: understanding how consumers act and why enabling predictions about their future behaviour. Predictive Analytics: extracting information from historical & real-time data to determine and predict future outcomes & trends.<br>**Contested issues:**<br>• Critical decisions based on risk scoring (using past data): loan and insurance applications, predicting criminal behavior, policing, etc.<br>• Targeted content & manipulation of consumer behavior (news, advertising, dating, etc.) | **Internet of Things (IoT)**<br>**Definition(s):** The inter-networking of 'smart' physical devices, vehicles, buildings, etc. that enable these objects to collect and exchange data. Capacity to do substantial analytics closer to the data source through Edge/Fog Computing.<br>**Contested issues:**<br>• Constant surveillance & ubiquitous sensors<br>• Machine-to-machine interactions with real life consequences<br>• Potential physical damages following cyber-attacks<br>• Processing at source & loss of information |
| **Blockchain DLT & Smart Contracts**<br>**Definition(s):** Technology underpinning digital currency, that secures, validates and processes transactional data. Distributed Ledgers: a trustless, consensus-seeking decentralized database with cryptographic sealing to secure data. Smart Contracts: the rules, possibly computer programmes, that codify transactions and contracts in line with underlying legal arrangements.<br>**Contested issues:**<br>• Security: possible identification of sensitive data<br>• Immutability to change<br>• Regulation in decentralized & permissionless DLT<br>• Laws & statutes as smart contracts?<br>• Environmental issues (electricity usage for 'mining' Bitcoin) | **Information Security (InfoSec)**<br>**Definition(s):** Preventing access, use, disclosure, disruption, modification, inspection, recording or destruction of information (both in physical and electronic terms).<br>Potential threats: cyber attacks, identity theft, intellectual property, computer malfunction, physical theft, etc.<br>**Contested issues:**<br>• Balanced protection: Confidentiality, Integrity, Availability (CIA) vs efficiency and productivity<br>• 'Acceptable' counter measures<br>• Information Security governance<br>• Personal data sharing<br>• Privileged vs distributed access to sensitive citizen data, etc. |

Each of the technologies listed are essentially linked to different stages of a typical software system design problem. While the Internet of Things is mainly a discussion of the automated high-quality data collection processes and the limits of processing data closer to the source, Information Security and Blockchain Distributed Ledger Technologies are mainly shaping the storage and access debates, also extending to the automation of the legal transactions controlled by *smart contracts*. In Big Data analytics, the main argument is the processing of huge amounts of highly varied and often unstructured data for practical purposes, such as finding patterns and correlations to obtain insight into user behaviour at present and to predict the future. With Artificial Intelligence, we have the fundamental scientific debates around new knowledge creation processes without fully understanding the process itself. The following three sections attempt to formulate a functional framework for the various ethical concerns in engineering and design terms.

## 3. Data Ethics

It is now a clear fact that we are becoming more and more predictable, and less and less discreet to private multinational companies, governments, individuals, and machines given the huge amounts of digital footprints we leave as a result of increasing online everyday activities and transactions. With the wider deployment of IoT systems, ubiquitous data collection will be taken to the next level with highly granular quality data becoming available through both human and machine activities. Although major recent data breaches, such as the Cambridge Analytica/Facebook case, and the introduction of

the European General Data Protection Regulation (GDPR), gave way to a growing public awareness around digital trust, privacy and data ownership/exploitation issues, the relevant technology and policy landscapes are still far from providing a clear vision.

**Research Directions and Engineering Challenges**

The range of ethical issues around 'data' can be broadly categorized into three stages as the *collection*, *storage*, and *access/sharing* processes:

i. At the *collection* stage questions should be asked around the choice of data (personal, public, government, commercial, etc.), the collection process (regular sampled vs. organic data, capturing and recording formats, quality of the data, standards and meta-data, etc.), and the assignment of the ownership of data (concerning design issues to create, access, modify, derive benefit from, sell, remove, assign responsibility, etc.).

ii. In relation to the *storage* of data, there are three major concepts: in centralized data infrastructures, one node essentially does everything; in distributed model, one data assigns roles for multiple sub-nodes; and in the fully decentralized model each node is connected to peers with no central node coordination.

iii. Issues concerning data *access/sharing* are harder to formulate: finding and locating relevant data is probably the first problem to deal with in the increasingly complex and wide-ranging data ecosystem. Identification of the ownership of data to permission access, privacy and security during data transaction and exchange, and the regulation of such processes and trust management appear to define the core engineering and design challenges.

Technology potentials are emerging mainly in two directions. First is to ensure privacy at the collection level through data minimization, anonymization, and encryption approaches – all coming with several disadvantages mainly due to loss of information and feature prioritization, and their vulnerability to cyber-attacks. Approaches to generate synthetic data aim to facilitate higher level processing on good quality data that highly resemble real datasets, with the advantage of avoiding privacy risks in particular. A second major technology direction is towards the generation of a global data infrastructure to support personal/private data and permissions. Blockchain distributed ledger and smart contract technologies in particular are emerging as the facilitators of such developments at the foundational level.

## 4. Algorithm Ethics

Individual choices, organisational operations and the public discourse are increasingly shaped by algorithmic decision systems as opposed to human self-determination. Examples range from customer decisions on online retail platforms (Netflix, Amazon, etc.), important life decisions such as loan/credit applications and recruitment decisions, or uses in policing and the justice system, to skewed realities through personalised content (e.g. web searches), and to the human actions based on inherently biased algorithmic systems. Most high-profile 'ethically problematic' incidents from recent discussions include the alleged voter manipulation in US elections, the use of risk assessment algorithms when sentencing criminals, inherent sexism observed through online translators, and racist tagging of images through popular online applications.

**Research Directions and Engineering Challenges**

The complexity needs to be broken into more manageable components in engineering terms: questions need to be asked in relation to the input data and the designer assumptions going into the system, problem formulation and definition of the outcomes, function of the human-in-the-loop, and the algorithmic process itself (rule-based based or black-box machine learning). Additionally, closer examination of the output functionalities, such as the various use cases of detection and recognition, decision support, 'intelligent' control and augmentation systems. Unintended outcomes in the form

of potential privacy breaches through integration of supposedly 'anonymised' data, or the discriminatory behaviour against protected attributes such as gender, race, age, etc. are amongst the areas of immediate public interest in relation to these discussions.

The two major technology challenges are around the elimination of *bias* and obtaining more *insight into the algorithm's behaviour*. There is clear overlap in these two objectives at an abstract level. Bias may be introduced into the system at various stages – it could be in the training data that amplified further in the process, or in the initial feature selection and the description of the problem and the desired outcome, or in the choice of the optimisation criterion. Current strategies to understand behaviour of the algorithm mainly progress in three directions [2]:

i. Outcomes-focused approaches forcing 'system accountability' providing limited insight into the process itself hence requiring constant training and monitoring;
ii. intervention-focused approaches where the algorithm is trained to also provide some 'rational' reasoning behind the outcome (e.g. "you did not qualify for the loan because you did not pay your last three rents on time but if you keep up with your payments in the next five months then you may be eligible … "); and,
iii. statistical analysis of the underlying model assumptions (more complex algorithms can unlock capabilities that simpler models cannot but at the cost of 'explainability').

Potential solutions include developing reference datasets and audit mechanisms to test bias and discriminative behaviour of algorithms, circuit breakers to ensure safety and security of algorithmic processing, stress testing of an algorithm's ethical performance and potential certification, incorporation of human oversight and feedback in the process, and further research into the transparency and explainability of the algorithmic design and execution processes.

## 5. Interaction Ethics

Increasing deployment of automated decision systems and interactive digital systems are giving way to new social spaces between humans and machines (e.g. social family robots), for which the traditional notions of 'ethics' are not sufficient to determine the right or wrong conduct. We now have conversational virtual assistants (Apple Siri, Amazon Alexa, etc.) that have capabilities to 'listen' to all everyday conversations and respond accordingly, which will potentially scale up to literally all devices with an on/off switch through the Internet of Things applications in the near future. There are also the machine-machine interactions with potential human costs, such as home appliances communicating with external systems (e.g. fridge communicating with supermarket). Face recognition technologies in everyday policing, targeted content and addictive behaviour design manipulating human decisions to take actions they would not otherwise do (social media, children content, online shopping/gaming platforms, etc.), machines replacing human tasks such as providing care for elderly or babysitting, machines becoming like an everyday partner or pet animal are all new interesting cases for ethical conceptualisation as well as for the technology design. There is also the evolutionary dimension, especially in the light of technological singularity discussions, but perhaps more urgently, in the cognitive shifts that are already visible, such as the drop in average attention span or the loss of ability to determine orientation without technology assistance due to extensive use of GPS systems, or the changing nature of human-space experience through the online communities and AR/VR systems.

The driverless car example is clearly an interesting one in illustrating some of the major challenges in this category. For example, in an unavoidable crash scenario, a critical ethical decision might be the prioritisation of the passenger's life against the maximum number of lives (pedestrians, passengers in the other car, cat on the road, etc.), although the ability for human drivers to make the most ethical decision in such situations is also questionable. Furthermore, the crash may be due to an algorithm anomaly or unpredictable human behaviour (driver in the other car), or a collusion might occur

between two driverless cars and both may appear to have acted 'properly'. There is also the obvious risk of potential hacks and the interactions with IoT infrastructure that will create a number of other ethical questions. All these concerns then raise questions around the way self-driving cars may be licensed (e.g. could driving tests be a solution?).

**Research Directions and Engineering Challenges**

Research into information security and cyber security is relatively well elaborated upon in the literature but topics around the 'acceptable behaviour' for human-machine and machine-machine interactions are almost non-existent (there are limited formalisation efforts through Asimov's 'three laws of robotics, 1950, and the EPSRC's principles of Robotics, 2010). For example, counter measures to control digital addiction or to distinguish between the legitimately immersive software from the 'exploitation-ware' is an immediate area of research interest. Similarly, prediction of both human and machine/software behaviour and their interdependencies are becoming increasingly interesting. Also, given the explosion of complexity, as more and more devices are introduced to the digital ecosystem, the automation of compliance and regulation also becomes a primary area of research.

## 6. Digital Systems and the Law

A key problem in the ever-growing digital ecosystem is the assignment of responsibility for the actions of algorithmic decisions [3]. A well-known example to illustrate this point is *Tay*, Microsoft's teenage bot, which started tweeting racist and sexist comments after a few hours of its launch, resulting in its complete shut-down within 24 hours. The statement from Peter Lee, the Corporate Vice President, read as follows:

> *"…we planned and implemented a lot of filtering and conducted extensive user studies with diverse user groups. We stress-tested Tay under a variety of conditions, specifically to make interacting with Tay a positive experience. Once we got comfortable with how Tay was interacting with users, we wanted to invite a broader group of people to engage with her…. Unfortunately, in the first 24 hours of coming online, a coordinated attack by a subset of people exploited a vulnerability in Tay. Although we had prepared for many types of abuses of the system, we had made a critical oversight for this specific attack..."*

Thus, it is hard to identify who is responsible for the wrong-doing of the algorithm with the reader being motivated to think that it was essentially a user misuse of the system in this case. However, at the other end of the spectrum, the success of DeepMind's *AlphaGo* has been almost entirely credited to the algorithm itself – not the programmers, manufacturers or the users.

Current discussion around the legal status of algorithms, no longer as solely intellectual property but rather recognition as 'artificial persons' similar to the status of companies in the law, is an interesting one. A further step is perhaps the recognition of robots as legal citizens (e.g. *Sophia*, the humanoid robot, was granted citizenship in Saudi Arabia in 2017). There is also an emerging debate around the 'rights' and 'obligations' of algorithms and robots: should robots only be 'slaves' to the humans? Could machine learning algorithms 'police' machine and human behavior autonomously? Can a robot autonomously 'decide' to harm or kill a human (e.g. war or terrorism case), etc.

## 7. Conclusions

Pandora's box is now open – digital systems are everywhere, having an impact on most of our everyday activities and critical decisions. Breaking up complex ethical problems in to components in engineering and design terms is an extremely difficult problem, hence the area is emerging as a distinct research field itself. It should be noted that most of the current discussions are around 'engineering ethics', which focuses on the behavior of the engineer rather than the behavior of the product, whereas this paper attempts to present an 'ethics engineering' concept to define research

directions and design processes to ensure digital systems are designed and regulated to behave ethically regardless of the application area or the use case.

## 8. Acknowledgements

I am extremely grateful to Philip Treleaven, David Hand and Emre Kazim for their comments and feedback on the first version of this paper. This work is also supported by EPSRC Impact Acceleration Account (IAA) award to UCL 2017-20 (Grant Reference: EP/R511638/1) and the EPSRC Next Stage Digital Economy Programme funding for the UK Regions Digital Research Facility (UK RDRF – Grant Reference: EP/M023583/1).

## 9. References

[1] Z. Engin and P. Treleaven, "Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies," *The Computer Journal,* p. bxy082, 2018.

[2] K. Hume, "When Is It Important for an Algorithm to Explain Itself?," *Harvard Business Review,* 06 July 2018.

[3] J. Barnett, A. S. Koshiyama and P. Treleaven, "Algorithms and the law," *Legal Futures,* 22 August 2017.