OF DUELS, TRIALS, AND SIMPLIFYING SYSTEMS

opinion on the AI HLEG "Policy and Investment Recommendations for Trustworthy AI" (2019)

PREPRINT

Giovanni Sileno
Systems and Networking Lab, Informatics Institute, University of Amsterdam
g,sileno@uva.nl

During medieval times, ordeals (*ordalia*) like judicial duels were rather formalised institutional constructs. When a dispute arose, contenders could decide or be pushed to put their fate under the 'judgement of God' by engaging in a duel. The mechanism was efficient in its simplicity: the process was well-established, rather rapid and the outcome was—to use a more modern term—*boolean*, leaving no space to ambiguity. Compare this with what occurred during the Roman era, in which magistrates were taught (and at least supposed) to construct their judgement by collecting and recording adequate evidence and testimonies, grounding their decision on interpretations of the law. This process was much slower and more expensive, the outcome less conclusive, as in certain conditions it could have been appealed, and it was expressed and often recorded in verbal forms. Indeed, then as now, *evidential* and *normative* reasoning were acknowledged to be rather complex endeavours. All these problems were quite rapidly solved with duels.

This (simplistic) historical reminder serves us to draw a fundamental analogy between human institutional systems and *artificial intelligence* (AI) systems. Resolving to settle a dispute through linguistic means, judicial duels can be put in correspondence to *machine learning* (ML) black boxes, typical artefacts of *sub-symbolic AI*. These devices receive rich inputs (e.g. a picture of an apple) and return synthetic outputs (e.g. an identifier to the apple category), via an opaque but direct quantitative process. In contrast, approaches based on explicitly declared evidence and a process of proof can be easily related to *automated reasoning* methods, a core topic of traditional *symbolic AI*. Applications based on symbolic AI typically receive input predicates in some formal language (e.g. round, sweet, red) and return as output some other predicate (e.g. apple) or a construct of those, through a traceable process of proof. As in trials, new (verbal) evidence or different applicable rules can produce a radically different outcome.

The sub-symbolic AI and symbolic AI approaches have coexisted with alternating fortunes in the field of artificial intelligence since its origin (and someone would add in philosophy, under the empiricist and rationalist flags), providing radically different solutions in terms of representations, processes, deployment, and impact. When sub-symbolic systems prove to function well in certain tasks (sometimes exceptionally well), this means that their tacit interpretative model is adequately aligned with the environment and the task in focus. Different to ordeals, ML-based systems have been constructed by means of an effective process of adaptation, put in place during a training phase, driven by rewards associated with what the designer indicated (directly or not) to be 'right'. But what is actually set as right? And can we be sure that this definition during the training phase is adequate to all possible uses of the system?

To further elaborate on this analogy, we need an additional step of abstraction. Ordeals, trials, as well as sub-symbolic and symbolic AI methods can be seen as manifestations of *simplification* processes, systematic patterns that emerge anywhere we see *systems* (biological, cognitive, linguistic, institutional and computational). As a matter of fact, any real system will experience moments of increased *complexity*—which, roughly, can be related to the number of *system failures* occurring while interacting with the *environment*, i.e. with what lies outside of the system's boundaries. Four scenarios are possible if complexity keeps increasing:

- the system simplifies the environment,
- the system finds a simplified (more efficient) informational niche sustaining positive interactions,[1]
- the system increases its internal complexity[2] (at higher costs),
- the system eventually collapses.

[1] See e.g. Heiner, R.: The origin of predictable behavior. *The American Economic Review*, 73(4), pp. 560–595 (1983).

[2] See e.g. the *law of requisite variety*, in Ashby, W. R.: Requisite Variety and Its Implications for the Control of Complex Systems. *Cybernetic*, (1:2), pp. 83–99 (1958).

Computational systems	*Institutional systems*
Machine learning	Ordeals
Automated reasoning	Institutional procedures
Humans in the loop (developers, users, ...)	Humans in the loop (legislators, judges, ...)

Table 1: Computational and institutional mechanisms offering different solutions to deal with complexity.

The second and third options concern primarily the informational dimension and form two poles of an imaginary axis of solutions to deal with complexity. On the 'simplifying' pole, ML methods reduce the richness of the original input to a small subset of features (which are found to be relevant to the task during the training phase), just as ordeals select a very limited number of elements to produce a definitive result. On the 'increasing internal complexity' pole, developers (sometimes system users) can introduce new exceptions to the system's behaviour in a structured or contingent way, just as legislators and to some extent judges and public officers can introduce refinements to the current normative legal system. Automated reasoning and formal institutional procedures can be seen as between these two extremes: in both cases, humans need to provide relevant explicit knowledge (evidence, theories and rules), adequately mapped to symbolic forms, and then processed in a transparent and traceable way.

For today's standards, ordeals are certainly not a reasonable form of legal judgment, just as reading omens is not a reasonable approach to medical diagnosis. It is therefore no coincidence that *explainable AI* and *trustworthy AI* emerged as relevant and urgent topics, particularly for applications in which human expertise already exists, strengthened by well-established processes of *justification*, against which we naturally confront AI performance. In light of the above, we can then say that *any explainable AI and trustworthy AI effort attempts to re-complexify computational systems that, as they are, lead to the risk of simplifying too much.*

* * *

This large preamble introduced the main ingredients of the arguments that this opinion paper advances. From a communication point of view, the *Policy and Investment Recommendations for Trustworthy AI*,[3] as well as the precedent *Ethics Guidelines for Trustworthy AI*,[4] provided by the high-level expert group on artificial intelligence set up by the European Commission, are meant to offer conceptual templates for justifying measures of actions at the individual and collective levels, aiming to positively guide the introduction and the impact of the pervasive use of AI.

Because any decision-making activity is based on assumptions about the *problem space*, it is tempting to use the documents to try to reconstruct the underlying assumptions, in terms of decision drivers and model of the world in which such directives are deemed to produce impact. Without the pretension of being exhaustive, this paper comments in particular on three aspects: the technical-legal dimensions of trustworthy AI (section 1), what we mean by AI (section 2) and the impact of AI (section 3).

1 On the technical-legal dimensions of Trustworthy AI

Whereas the 'Ethics Guidelines for Trustworthy AI' focuses on identifying general ethical principles for the development, deployment and use of AI systems, together with indications on how to realise such principles and assess their operationalisation, the 'Policy and Investment Recommendations' document pays attention to how to promote research, development and use of trustworthy AI in the current European socio-economic context. Intuitively, the former document is more concerned by *drivers*, while the latter relies more on a *world model*.

From further inspection, a seemingly slight shift already occurs at the very beginning of the two documents. Trustworthy AI is defined in both along three dimensions: *lawful*, *ethical* and *robust*. However, the guidelines explicitly focus on only the second and third aspects, whereas the policy recommendations implicitly cover the whole. This subtle difference arouses suspicion: *Is the legal component deemed to be ancillary in the constitution of trustworthy AI?*

Indeed, in the guidelines document (section 2.1), the rule of law is explicitly called on only with respect to the use of *privacy-by-design* and *security-by-design* conception methods. These methods are not *constructed* by law, but the law (is deemed to) impose their application during the design of computational systems. Therefore, in essence, this view is similar to the one used in domains such as manufacturing: the law sets certain standards of security to be followed

[3] https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

[4] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

(e.g. protocols and requirements) and manufacturers implement them (in their processes and products) in order to be compliant. If this observation is valid, it also hints to potential limitations.

Regulation typically requires to have some knowledge about the behaviour that is going to be regulated. If, for instance, we are talking about regulating agriculture, we know to a good extent the type of conduct that we can expect (from the relevant socio-economic actors, given the available knowledge and technology) and their impact. Instead, it is more complicated to talk about regulating computational technology, because technological innovation here modifies, at a very fast pace, our interfaces with the world, and it continuously introduces new *affordances* with an overall impact that is often not clear during the design phase.[5]

A general point underpinning many of the arguments in the two documents—as well as in most of the contemporary projects in *ethical AI*, *responsible AI* and related tracks—concerns the importance of performing an analysis of the consequences of system deployment as completely as possible at the design-time while maintaining adequate reassessment during run-time.

To realise this, the typical solution consists of a panel of experts (in ethics, law, etc., and stakeholders) on the one hand, while on the other hand AI and IT developers apply the outcomes of the panel into their implementations. Even the policy recommendations on education strengthen this *panel-solution* idea: to sustain such a vision we clearly need experts, on both sides, and we need therefore to allocate adequate educational resources to realise this.

Rather than a disruptive vision, however, this seems to be a rather traditional approach, and certainly not a *scalable* one. From what we know about the interactions between policy, legal and IT departments in administrative organisations, and also in business-IT alignment contexts, a too clear-cut decomposition of concerns naturally creates frictions due to *mutual misalignments*.[6]

Let us consider the general division in organisations between: (a) *policy*, (b) *design/development* and (c) *operations* spheres of activities. Typically, people at the design/development level (and even less at the policy level) are not synchronised with the problems observed at the operations level, while people at the operations level (e.g. customer services) usually do not have the power to modify the functioning of the system, even when they recognise that the case they have in front of them is not treated properly. In this context, failures accumulate, and for economic reasons, only the most frequent or critical types of failures are eventually treated, while the rest remain unsolved in a long tail of ‘unfortunate’ cases. These mechanisms of *bureaucratic alienation*—which all of us experience at some point when dealing with public or private organisations—can be reduced only if the design-iteration cycles are essentially continuous, or, even more radically, completely bypassed.

This would mean that the same panel of experts (and stakeholders), rather than producing a linguistic artefact giving requirements to be reinterpreted by developers embedded in their own conceptual niche, take *direct responsibility* in the development of the system. This vision can be restated as *Nobody better than who sees the wrong (right) can describe what is wrong (right), and these descriptions should be the drivers to which the system should ‘intelligently’ adapt*. Then, the role of AI and IT researchers/developers would not be one of engineering and maintaining *ad-hoc* solutions, but of creating adequate interfaces and computational intelligence for this systematic requirement.

To develop this idea further, consider how private law and functionally similar legislation enable people to create *ad-hoc* normative mechanisms under certain conditions, which, if valid, are accounted for and protected by the legal system. Now suppose that members of an association wanted to introduce the following paradoxical norm in their statute: by saying ‘CIAO’ to someone, a member would become a slave of the addressee of the greetings (which, by the way, is the etymological origin of the Venetian word *sciao*, although in the sense of ‘I am your servant’). This is not possible in our society, as legislators and jurisprudence have set adequate constraints: higher-order norms would *invalidate* such an institutional construct, and, if the outcome were to occur in any case (someone becomes *de facto* enslaved), enforcement measures would be activated by competent authorities. In other terms, law offers an example of what we may call *deferred conception*: it is used to circumscribe future (lower-order) norms, even if we do not know at the moment how they will be or might look like.[7]

[5] The concept of affordance was introduced in ecological psychology (Gibson, J.: *The ecological approach to visual perception*. Houghton Mifflin [1979]) and gained a primary role in the design of objects and interfaces: if an object is meant to be a glass, we need to perceive its ‘drinking’ affordance, i.e. that the object affords our drinking. The concept is applicable also to communicative actions, including institutional actions.

[6] See for instance Seddon, J.: *Systems Thinking in the Public Sector.* Triarchy Press (2008); Benbya, H. & McKelvey, B.: Using Coevolutionary and Complexity Theories to Improve IS Alignment: A Multi-Level Approach. *Journal of Information Technology*, 21, pp. 284–298 (2006).

[7] See e.g. the discussion on permissive norms, as for instance Makinson, D.: On a fundamental problem of deontic logic. *Norms, Logics and Information Systems*, pp. 29–53 (1999), and Alchourrón, C. E., & Bulygin, E.: Pragmatic foundations for a logic of

The title of this section refers to the 'technical-legal' dimension because what we miss today is an analogous 'deferred conception' level for computational development, which could be named *normware*, to be distinguished from software and hardware. In recent work,[8] we argued that normware is functionally expressed by two dimensions. On the one hand, normware consists of computational artefacts specifying norms, i.e. providing the designer with normative categories to define, qualify and circumscribe behaviour, possibly described at a higher level of abstraction than the operational level. Note that with respect to this dimension, software could be interpreted as a subset of normware by reading instructions as commands. However, it is on the second dimension that we can observe a neat distinction. A piece of normware is not (and should not be) used for control, only for guidance. It is meant to sustain specific coordination mechanisms for autonomous computational components, and typically it is part of an *ecology* of possibly conflicting normative components. Only under this assumption would we be able to maintain a pluralism similar to the one occurring in human societies, giving space to private parties, intermediate bodies and public authorities to operate in a modular way on the system.

The fact that we do not have a normware level of conception for artificial devices today does not imply that it cannot be. On the contrary, we know from observing maintenance processes in human social systems that, from a functional point of view, it *has* to be introduced.

To conclude, legal systems are certainly regulative, but they also provide affordances (e.g. of creating agreements) for creating affordances (enabling actions otherwise impossible), albeit regulated.[9] This 'enabling' or 'empowering' dimension of law, as well as an automated/embedded view on regulatory processes within computational systems, is completely unexplored in the two documents. This second oversight is particularly unexpected: regulatory technologies are a long-running research field, partially overlapping with the traditional AI & Law subfield,[10] and have recently resurged rebranded as RegTech—although in much more simplified forms.

2 On what we mean by AI

Strangely enough, the recommendations and guidelines documents do not explicitly mention what is precisely meant by AI. Many passages, such as the reference to the construction of a data-sharing and computing infrastructure for AI at the European level, suggest that the term is mostly used in the conflated, contemporary meaning of sub-symbolic AI. This simplification needs to be contested for several reasons.

First, AI is not just machine learning. As suggested in the introduction, many of the problems (explainable AI, trustworthy AI, ...) that this family of approaches have can be plausibly solved only by attempting to integrate sub-symbolic processing with some symbolic-level aspect. If this hypothesis is true, limiting fundamental research and education on symbolic AI or other related tracks, as well as the many disciplines which give models and inspirations to these contributions, is a strategic error. If we are pursuing rationality (rational systems, rational institutions, etc.), it is rather implausible that this will be obtained only by empirical means.

Second, AI as a discipline took a pragmatic approach towards the definition of intelligence: intelligence is *appropriate behaviour*. For this reason, the term AI could map essentially *to any informational processing system*, as such a definition is independent of whether the system is realised via bits in a computer or communications in a social system.[11] As a matter of fact, any institutional system is *super-human* in the sense that it transcends the individuals that compose it, and, up to the extent which such a system is designed—i.e. when its form has been deliberately chosen by rational agents—it can be seen as an artificial one.

This conceptualisation implies that the proposed recommendations can in principle be transposed to public and private organisations. But, as a provocation, can we imagine ethical committees judging whether a company behaves in an ethical way, or how it should behave? Existing examples, such as ethical committees for medical research, operate in narrow contexts and typically decide on issues about which law is still silent. The pervasive use of AI, by contrast, has the potential to impact all human activities, many of which are already regulated by law, at least with respect to

norms. *Rechtstheorie*, 15, pp. 453–464 (1984). These authors observe that explicit permission is needed in real-life normative systems to limit the authority of subordinate instances to create new norms against it.

[8] Sileno, G., Boer, A., & van Engers, T.: The Role of Normware in Trustworthy and Explainable AI. *Proceedings of the XAILA workshop on eXplainable AI and Law*, in conjunction with JURIX 2018 (2018).

[9] Without forgetting that law in itself, in order to operate, builds upon other affordances, such as e.g. those enabled by the printing press, see Hildebrandt, M.: *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*, Edward Elgar Publishing (2015).

[10] For a technical overview of the field of AI & Law, see Bench-Capon, T., Araszkiewicz, M., Ashley, K. et al., A history of AI and Law in 50 papers. *Artificial Intelligence and Law*, 20:215 (2012).

[11] Along similar lines see e.g. Bryson, J. J., & Theodorou, A.: How Society Can Maintain Human-Centric Artificial Intelligence. *Human-Centered Digitalization and Services*, pp. 305–323 (2019).

undesired outcomes. Indeed, legal experts can make a stronger for having a voice in this than ethical committees. However, as I argued in the previous section, if we cannot go beyond the panel paradigm for design/maintenance, then we are opening doors to bureaucratic alienation, which in this case will be replicated *for each* AI system.

Third, rather than pushing students towards computational-AI specific education, a more sound option would be to accept that AI covers much more than the computational aspects of AI and enrich existing disciplines with research tracks attempting to understand the applications of available *commoditised* computational approaches in integration with the existing corpus of knowledge. This would favour cross-fertilisation and hybridisation, but also the maintenance of research tracks, methods and paradigms that have been already established and streamlined across generations, and that form unique and irreducible signatures.

3 On the impact of AI

At a more general level, all AI techniques build upon some form of optimisation. Successful optimisation brings increasing returns, and the overall coupling of a *system* with an *environment* typically exhibits *value extraction* patterns. This optimisation is not without negative effects, and just as mining produces pollution, the environmental, individual and societal costs of AI typically go along with generating increasing returns with respect to some task. It is well known that the image of AI that the media conveys overlooks the human costs required for its development (e.g. the millions or billions of users interacting on platforms[12]) and deployment (e.g. operators more or less constrained to following AI-system indications, limiting human autonomy).

In the previous section we started to argue that, from a strategic point of view, there are reasons to question the excessive focus on sub-symbolic AI research and development suggested by the policy recommendations on trustworthy AI. Here we add to this considering three aspects: its environmental impact; the risk of becoming an obsolete technology (at least with respect to its current form); and the risk of becoming a 'transparent' technology, introducing invisible cognitive and societal dependencies.

Machine learning is a technology that literally pollutes. Recent estimates in natural language processing suggest that the carbon footprint of training a single AI system is around five times the lifetime emissions of an average car.[13] Indeed, each training involves an adaptation comparable in quality to the evolutionary adaptation of a natural species. However, both nature (through selection and reproduction) and nurture (through education) have found mechanisms for 'reuse' that ML today does not have (or at least, only up to a certain point). It mostly relies on brute force approaches with plenty of data and lots of computing power. There is, therefore, a risk in investing in huge infrastructures for data maintenance and processing knowing that at some point a 'killer' theoretical advance could (and should) make it all obsolete.

A similar consideration could be made about education. If interfaces between humans and computers continue to improve, the need for 'technical' computational expertise will in principle decrease, save for a core number of people required to maintain and possibly improve the infrastructure. Furthermore, the kind of non-verbal, intuitive knowledge necessary to select and tinker with ML methods today is the typical domain in which ML methods might be particularly effective (if we solve the problem of eagerness of data). There is therefore the risk of forming a generation of the work-force whose appeal could soon be diminished. In contrast, competence in human matters might become a more valuable asset in a computational-entrenched society, as it might counterbalance gaps resulting from the pervasive use of artificial systems.[14]

Recommendation 5.1 is particularly relevant in this respect. It is plausible that the development of cognitive skills in humans will be affected, and we need to be conscious about which tasks will be affected, to what degree, and whether this change is acceptable for the idea that we want to maintain of us as humans. In contrast, it is arguable that 'education and skills [..] are essential in a world where "intelligent" systems perform an increasing number and variety of tasks' (ch. 1, D). If systems become sufficiently complex to be able to adapt and repair themselves without our intervention, it is plausible that humans will be pushed to simplify their conceptualisation to *animist*-like positions. It is still an open question whether, in order to give individuals the opportunity to flourish with respect to their psycho-physical constraints, a bar should be set on the pervasiveness of automation into the human existential experience.

[12] See e.g. Casilli, A., & Posada, J.: The Platformization of Labor and Society. *Society and the Internet. How Networks of Information and Communication are Changing Our Lives*, pp. 293–306. Oxford University Press. (2019)

[13] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. available at: http://arxiv.org/abs/1906.02243

[14] Maintaining non-computationally driven procedural knowledge is also critical in preparing for any event in which technology might suddenly stop working.

Finally, having reached human-cognitive aspects, it is also interesting to look at the opposite pole of the technological chain: the production of electronic equipments. The policy recommendations do not manifestly contain any reference to this, but, the more our society becomes dependent on technology, the more Europe will be geo-politically *locked-in* by who is providing us technology, exposing us to security problems for which no technological solution exists. There is no such a thing as a trustless technology.[15]

Conclusions

The introduction of ubiquitous cyber-physical connections in all human activities raises serious concerns at the societal, cognitive, and environmental levels, and their potential impact is too critical to be belittled by the belief in a technologically-driven 'magnificent and progressive fate'. Therefore, all initiatives contributing to the discussion on how to guide future technological transitions are welcome, and are particularly important when reaching higher institutional levels, where the allocation of public resources is decided. However, it is also essential that these initiatives care, in particular at this higher level, about pluralism and diversity with respect to the sensibilities and expertise of all disciplines. This concern is a strategic one.

The *Policy and Investment Recommendations for Trustworthy AI*, provided by the high-level expert group on artificial intelligence set up by the European Commission, and the previous *Ethics Guidelines for Trustworthy AI*, are undoubtedly important steps in publicly setting the agenda for advancing a common societal strategy on these topics. Reading the two documents, one can see the effort to provide concrete suggestions that are applicable in the short term, and one can appreciate the organised structure and overall internal consistency. However, as this text attempted to convey, gaps can be observed in the underlying assumptions, some of which cannot be easily dismissed. The oversimplifications concern mostly how AI is defined (sub-symbolic systems instead of general informational processing systems), the interface between AI and institutions (neatly separated instead of continuity), and the plausible evolution at the technological level (expecting a plateau instead of potentially near disruptive innovation).

Ultimately, I believe that the passage from 'duel'-like to more 'trial'-like mechanisms for computational decision-making needs urgently to be acknowledged as a fundamental requirement for trustworthy AI, which in turn demands technologies enabling a computational 'jurisprudence', including mechanisms similar to private law. The continuity between AI and institutions does not imply that such computational jurisprudence will remove the need for a human one. Their relationship should be rather similar to that of higher-courts towards lower-courts, ensuring that humans will always remain in control and mechanisms like *appeals* and *cassation* become systemic.

In this light, the recommendations of the high-level expert group can perhaps better be seen as *tactical* directives. The proposed investments towards ML (general education, infrastructures, etc.) will certainly facilitate the introduction of services or applications relying on big-data in the private and public sectors, which are plausibly necessary to maintain competitiveness at global level—although on a rather *follower* attitude. However, this reading means that strategic issues still need to be considered, such as maintaining relevant research/education tracks, even if not directly related to machine-learning applications, as well as counterbalancing technological dependence. This equally requires an adequate allocation of resources.

[15] See the classic Thompson, K.: Reflections on trusting trust. *Communications of the ACM*, 27(8), pp. 761–763. (1984)