Of duels, trials and simplifying systems

Sileno, G.

DOI
10.1017/err.2020.38
Publication date
2020
Document Version
Final published version
Published in
European Journal of Risk Regulation
License
CC BY

Of Duels, Trials and Simplifying Systems

Giovanni SILENO*

This short paper aims to unpack some of the assumptions underlying the "Policy and Investment Recommendation for Trustworthy AI" provided by the High-Level Expert Group on Artificial Intelligence (AI) appointed by the European Commission. It elaborates in particular on three aspects: on the technical-legal dimensions of trustworthy AI; on what we mean by AI; and on the impact of AI. The consequent analysis results in the identification, amongst others, of three recurrent simplifications, respectively concerning the definition of AI (sub-symbolic systems instead of "intelligent" informational processing systems), the interface between AI and institutions (neatly separated instead of continuity) and a plausible technological evolution (expecting a plateau instead of a potentially near-disruptive innovation).

I. Introduction

From a communication point of view, the "Policy and Investment Recommendations for Trustworthy AI",[1] as well as the precedent "Ethics Guidelines for Trustworthy AI",[2] provided by the High-Level Expert Group on Artificial Intelligence (AI) set up by the European Commission, offer conceptual templates for justifying measures of actions to private and public actors, meant to positively guide the impact of a pervasive use of AI. Unpacking the implicit assumptions upon which these documents build is therefore a useful exercise for evaluating the soundness and strength of their propositions. Without the pretension of being exhaustive, this short opinion paper will comment in particular on three aspects that are addressed to a limited extent or surprisingly overlooked in the two documents: the technical-legal dimensions of trustworthy AI (Section II); what we mean by AI (Section III); and the impacts of AI (Section IV). For this purpose, the following introductory section will provide a few conceptual ingredients for the arguments advanced here, by means of (simplistic) historical examples, drawing a fundamental analogy between human institutional systems and AI systems.

* Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands; email: g.sileno@uva.nl. This work was partly funded by NWO (VWData project).

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

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

European Journal of Risk Regulation, 11 (2020), pp. 683–692 doi:10.1017/err.2020.38

1. Of duels and trials

During mediaeval times, ordeals (*ordalia*) like judicial duels were rather formalised institutional constructs. When a dispute arose, contenders could decide or be pushed to leave their fate to the "judgment 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 for ambiguity. Compare this with what occurred during the Roman era, in which magistrates were taught (or at least supposed) to construct their judgments 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 of these problems were quite rapidly solved with duels.

2. Of symbolic and sub-symbolic AI

Renouncing to settle a dispute through linguistic means, judicial duels can be put in relation to *machine learning* (ML) black boxes, typical artefacts of *sub-symbolic AI*. These devices receive rich inputs (eg a picture of an apple) and return synthetic outputs (eg 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 (eg "round", "sweet", "red") and return as output some other predicate (eg "apple"), or a construct of those, through a traceable process of proof utilising knowledge expressed in computational forms. As in trials, new (verbal) evidence or different applicable rules can produce radically different outcomes.

The sub-symbolic AI and symbolic AI approaches have coexisted with alternating fortunes in the field of AI 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. Differently from 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?

3. Of systems that simplify

Ordeals, trials and 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,

Table 1. Analogies between informational processing constructs in computational and institutional systems.

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

any real system will experience moments of increases of *complexity* – which, roughly, can be related to the number of *system failures* occurring while interacting with the *environment* (ie with what lies outside of the system's boundaries).

Four scenarios are possible if complexity keeps increasing. Either:

- The system simplifies the environment;
- The system finds a simplified (more efficient) informational niche sustaining positive interactions;[3]
- The system increases its internal complexity[4] (at higher costs); or
- The system eventually collapses.

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 being 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 (Table 1).

4. The roots of explainable AI and trustworthy AI

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 (healthcare, legal services), strengthened by well-established processes of *justification*

[3] See eg R Heiner, "The origin of predictable behavior" (1983) 73(4) The American Economic Review 560.

[4] See the *law of requisite variety*, eg in WR Ashby, "Requisite Variety and Its Implications for the Control of Complex Systems" (1958) 1(2) Cybernetica 83.

against which we naturally confront AI performance. In light of the above, we can then say that any explainable AI and trustworthy AI efforts attempt to *re-complexify* computational systems that, as they are, lead to the risk of simplifying too much.

II. 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.

From further inspection, a seemingly slight shift 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 imposes (or is deemed to) 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 (eg 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 at potential limitations.

Regulation typically requires having some knowledge about the behaviour that is going to be regulated. If, for instance, we are talking about regulating agriculture, we know to a large extent the types 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 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 design-time while maintaining adequate reassessment during run-time.

[5] The concept of affordance was introduced in ecological psychology (J Gibson, *The ecological approach to visual perception* (Boston, MA, 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 (ie that the object affords our drinking). The concept is, however, also applicable to communicative actions, such as in institutional actions.

To realise this, the typical solution consists of a panel of experts (in ethics, law, etc., and stakeholders) on the one hand and AI and information technology (IT) developers on the other applying the outcomes of the panel to 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 therefore need to allocate adequate educational resources in order 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, but also in business–IT alignment contexts, a too clear-cut decomposition of concerns naturally creates frictions due to *mutual misalignments*.[6] As an illustrative example, let us consider the general division in organisations between (1) *policy*, (2) *design/development* and (3) *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 (eg 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 being treated properly. In this context, failures accumulate, and for economic reasons, only the most frequent or critical types of failures will eventually be treated, while the rest will continue to stay 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 linguistic artefacts that need to be reinterpreted by developers embedded in their own conceptual niches, would take direct responsibility in the very development of the system. This vision can be re-instantiated 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 greeting (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

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

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 how they might look.[7]

The title of this section refers to a "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 (ie providing the designer with normative categories to define, qualify and circumscribe behaviour, possibly described at a higher level of abstraction than the operational level).[9] With respect to this first dimension, software could be interpreted as a subset of normware by reading instructions as commands, although concepts as prohibition do not have a direct procedural characterisation. 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 of the computational system.[10] 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.[11] 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.[12] The fact that we do not have a normware level of conception for artificial devices today does not imply that this cannot be achieved. On the contrary, we know

[7] See the discussion on permissive norms, eg D Makinson, "On a fundamental problem of deontic logic" in P McNamara and H Prakken, *Norms, Logics and Information Systems* (Amsterdam, IOS Press 1999) pp 29–53; CE Alchourrón and E Bulygin, "Pragmatic foundations for a logic of norms" In *Rechtstheorie 15* (Berlin, Dunker & Humblot 1984) pp 453–64. 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] G Sileno, A Boer and T van Engers, "The Role of Normware in Trustworthy and Explainable AI" In *Proceedings of the XAILA workshop on eXplainable AI and Law, in conjunction with JURIX 2018*, CEUR WS Proceedings, Vol. 2381 (2018) pp 9–16.

[9] This definition aligns with the matter studied in the "normative systems" research field, gathering contributions belonging to philosophical logic, legal philosophy and computer science. See, eg, AJI Jones and M Sergot, "On the Characterisation of Law and Computer Systems: The Normative Systems Perspective" in J-JC Meyer and R Wieringa, *Deontic Logic in Computer Science: Normative System Specification* (Hoboken, NJ, J. Wiley 1993) pp 275–307.

[10] "Control" is here used in the computational sense: the *control flow* is the order of instructions, and a *controller* is a component directing certain computational operations. In contrast, "guidance" is meant to capture that the regulated components are autonomous and might not follow what the controller requires.

[11] In addition to the normative system field, problems of aggregation of directives and of other preferential structures are studied extensively in computational social choice; see, eg, F Brandt, V Conitzer, U Endriss, J Lang and AD Procaccia, *Handbook of Computational Social Choice* (Cambridge, Cambridge University Press 2016). The ecological perspective is investigated in fields such as multi-agent systems and agent-based modelling and simulation. At the moment, however, none of these fields has had a visible impact in software engineering practices.

[12] Consider, for instance, a research institute that maintains a registry of patients with a rare disease. In principle, access to a patient's data should be granted only by taking into account the interpretation of the patient's consent, of the specific data-sharing agreement held with the requestor and of relevant legislation, such as the General Data Protection Regulation (GDPR). In practice, all of these directives (or usually a simplified, conservative version of these) are today manually encoded into a single computational policy enforced by the access control module, thus limiting transparency and actionability, if not the rights of the parties involved.

from observing maintenance processes in human social systems that, from a functional point of view, this has to be introduced.

To conclude, legal systems are certainly regulative, but they also provide affordances (eg of creating agreements) for creating affordances (enabling actions otherwise impossible), *albeit* regulated.[13] This "enabling" or "empowering" dimension of law, as well as an automated/embedded view of regulatory processes within computational systems, is completely unexplored in the two documents. This second oversight is particularly unexpected: regulatory technologies represent a long-running research field, partially overlapping with the traditional AI and law subfield,[14] and they have recently resurged rebranded as *RegTech* – although in much more simplified forms.

III. 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 ML. As suggested in the introduction, many of the problems (explainable AI, trustworthy AI, etc.) 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 that 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 by empirical means only.

Second, AI as a discipline took a pragmatic approach towards the definition of intelligence: intelligence is *appropriate behaviour*.[15] 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.[16] As a matter of fact, any institutional system is

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

[14] For a technical overview of the field of AI and law, see T Bench-Capon et al, "A History of AI and Law in 50 papers: 25 Years of the International Conference on AI and Law" (2012) 20 Artificial Intelligence and Law 215.

[15] Definitions of AI to be found in the literature: "... doing well at a broad range of tasks is an empirical definition of 'intelligence'", H Masum (2002); "Intelligence is the computational part of the ability to achieve goals in the world. Varying kinds and degrees of intelligence occur in people, many animals and some machines", J McCarthy (2004); "Any system ... that generates adaptive behaviour to meet goals in a range of environments can be said to be intelligent", D Fogel (1995); "... the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral subgoals that support the system's ultimate goal", JS Albus (1991). For source references and further definitions, see S Legg and M Hutter, "Universal intelligence: A definition of machine intelligence" (2007) 17(4) Minds and Machines 391.

[16] On similar lines, see, eg, JJ Bryson and A Theodorou, "How Society Can Maintain Human-Centric Artificial Intelligence" in M Toivonen and E Saari *Human-Centered Digitalization and Services* (Amsterdam, Springer 2019) pp 305–23.

super-human in the sense that it transcends the individuals that compose it, and, to the extent that such a system is designed (ie 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 on highly contingent issues. 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 undesired outcomes. Indeed, legal experts can make a stronger case 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. This means enriching existing disciplines with research tracks that integrate their existing corpus of knowledge with applications of available, commoditised computational approaches. 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.

IV. 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 system with environment typically exhibits *value extraction* patterns. This optimisation is not without negative effects, and just as mining produces pollution, generating increasing returns through AI with respect to some task typically goes along with environmental, individual and societal costs. It is well known that the image of AI that the media conveys overlooks the human costs required for its development (eg the millions or billions of users interacting on platforms[17]) and deployment (eg 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 by considering three aspects: AI's environmental impact; the risk of it becoming an obsolete technology (at least with respect to its current form); and the risk of it becoming a "transparent" technology, introducing invisible cognitive and societal dependencies.

[17] See, eg, A Casilli and J Posada, "The Platformization of Labor and Society" in M Graham and WH Dutton, *Society and the Internet: How Networks of Information and Communication are Changing Our Lives* (Oxford, Oxford University Press 2019) pp 293–306.

ML 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.[18] 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, has only up to a certain point). ML 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 reduce the requirements of the amount of data necessary for such learning). There is therefore the risk of educating a workforce that will soon experience much diminished appeal for its skills. In contrast, competence in human matters might become a more valuable asset in a computationally entrenched society, as it might counterbalance gaps resulting from the pervasive use of artificial systems. Maintaining non-computationally driven procedural knowledge is also highly critical to preparing for any event in which technology might suddenly stop working.

Recommendation 5.1 of "Policy and Investment Recommendations for Trustworthy AI" 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 to the idea of ourselves as humans that we want to maintain. 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" (chapter 1, D, "Policy and Investment Recommendations for Trustworthy AI"). 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 conceptualisations to *animist*-like positions. It is still an open question as to 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 in human existential experience.

Finally, having reached human cognitive aspects, it is also interesting to look at the opposite pole of the technological chain: the production of electronic equipment. 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 geopolitically locked in by who is providing us with that technology, exposing us to security problems for which no technological solution exists. There is no such a thing as a trustless technology.[19]

[18] E Strubell, A Ganesh and A McCallum, "Energy and Policy Considerations for Deep Learning in NLP" (2019). Available at: <http://arxiv.org/abs/1906.02243>.

[19] See the classic K Thompson, "Reflections on trusting trust" (1984) 27(8) Communications of the ACM 761.

V. 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 they are particularly important when such transitions might reach higher institutional levels, where the allocation of public resources is decided. However, it is also essential that these initiatives care, particularly 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 AI 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 paper has 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 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 human jurisprudence. Their relationship should be rather similar to that of higher courts towards lower courts, ensuring that humans will always remain in control and mechanisms such as *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 the global level – although with a "follower" attitude with respect to the other actors. 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 ML applications, as well as counterbalancing technological dependence. This equally requires an adequate allocation of resources.