Technology and Regulation

Trustworthy AI
a cooperative approach

Jacob Livingston Slosser*; Birgit Aasa**; Henrik Palmer Olsen***

Artificial Intelligence; Trustworthy AI; Algorithmic Decision Making; Administrative Law

jacob.slosser@jur.ku.dk
birgit.aasa@alumi.eui.eu
henrik@jur.ku.dk

The EU has proposed harmonized rules on artificial intelligence (AI Act) and a directive on adapting non-contractual civil liability rules to AI (AI liability directive) due to increased demand for trustworthy AI. However, the concept of trustworthy AI is unspecific, covering various desired characteristics such as safety, transparency, and accountability. Trustworthiness requires a specific contextual setting that involves human interaction with AI technology, and simply involving humans in decision processes does not guarantee trustworthy outcomes. In this paper, the authors argue for an informed notion of what is meant for a system to be trustworthy and examine the concept of trust, highlighting its reliance on a specific relationship between humans that cannot be strictly transmuted into a relationship between humans and machines. They outline a trust-based model for a cooperative approach to AI and provide an example of what that might look like.

1. Introduction: The concerns about Algorithmic Decision-Making systems

The literature on trustworthy AI is booming. A quick Google Scholar search for "trustworthy AI" gives more than 11,000 hits for 2022 alone with the total number in the past five years is around 24,400. This is no surprise given the rise of algorithmic decision making systems (ADMs) whose flaws are well documented and have raised numerous concerns in the fields of human rights and social justice.[1] Numerous academics and investigative reporters have demonstrated how the use of algorithms in social welfare are far from trivial.[2] These systems require serious attention before, during, and after deployment, particularly those that are procured by public institutions. Still, ADM technology offers significant potential to greatly expand rights protections. Public institutions have a duty to ensure the rights of their citizenry not only by protecting from harm, but by guaranteeing the equal application of law across a populace. Decision making processes in public administration, though based on the same laws, can vary significantly among the many decision makers within their employ. Often, individual case workers and administrators spread throughout a jurisdiction decide on similar cases without ever knowing the particulars of how a law has been applied in those other circumstances.[3] As ADM tech becomes ever more capable and the economic burden of running an increasingly complex public administration grows, the push for using ADM is likely to continue, if not increase. The balance is delicate; the solutions, not so straightforward.

Far and away the leading candidate for solutions are the use of ethical guidelines for those designing and implementing these systems,[4] leading at least in terms of its ubiquity. This is not to discount recent legislative attempts at incorporating ethical principles into more binding forms.[5] Regardless, in both formats ADM systems and human

1 For a brief overview of the many concerns raised, see the Written Testimony of Meredith Whittaker Co-founder and Co-director, AI Now Institute, New York University to the United States House of Representatives Committee on Science, Space, and Technology of June 26, 2019, available at: https://ainowinstitute.org/062619-whittaker-house-testimony.pdf. See also, the now ubiquitous: Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books 2017).

2 See Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press 2018).

3 Ann Light and Anna Seravalli, 'The Breakdown of the Municipality as Caring Platform: Lessons for Co-Design and Co-Learning in the Age of Platform Capitalism' (2019) 15 CoDesign 192.

4 For review see, Brent Mittelstadt, 'Principles Alone Cannot Guarantee Ethical AI' (2019) 1 *Nature Machine Intelligence* 501; Emre Kazim and Adriano Soares Koshiyama, 'A High-Level Overview of AI Ethics' (2021) 2 *Patterns* 100314.

5 Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts 2021.Doc nr.

* Jacob Livingston Slosser, Assistant Professor, University of Copenhagen, Faculty of Law, CECS – Center for European and Comparative Legal Studies.
** Birgit Aasa, Lawyer Linguist, European Parliament, Brussels, This article expresses the author's personal views only and cannot be attributed to the European Parliament.
*** Henrik Palmer Olsen, Professor of Jurisprudence, University of Copenhagen, Faculty of Law, iCourts - the Danish National Research Foundation's Centre of Excellence for International Courts; Research fellow 2023-2024 at IEA-Paris.

Received 21 Aug 2023, Accepted 24 Sep 2023, Published 25 Oct 2023

decision makers are seen as a world apart with the dangers presented by the former being safeguarded by the latter. Not so much a balance to be struck, but a harm to be avoided. A common safeguard in this framing has been to invoke concepts traditionally embedded within human institutions, like trustworthiness. Yet, not much attention has been given to unpacking the concept or defining what it requires. One oft-proposed suggestion is to ensure a trustworthy ecosystem via human proxy: a human-in-the-loop (HITL). However, it is not clear where that human should be placed within the loop, or what exactly they should do once they are there.[6]

Take for example the proposed Artificial Intelligence Act (AIA).[7] Among the prescribed requirements to ensure that AI remains trustworthy is Art. 14 concerning "human oversight"[8]. This article requires that AI systems: can be "overseen by natural persons"; has human oversight built into the system with the aim of "preventing or minimising the risks to health, safety or fundamental rights"; are built so that natural persons can "understand [its] capacities and limitations"; support awareness of possible over-reliance on the system (automation bias); can be understood and interpretable by natural persons; enable override by natural persons; and, are capable of being turned off. What this amounts to, is a requirement that natural persons must be able to understand and interfere with the functioning of high-risk AI systems. This approach buries the definition of trustworthy AI into elements that rather than clarify, add a spate of further questions: e.g. when is an AI system understandable? When is an output "interpretable" and when should a natural person "override" an AI system? These (among many other questions) remain unexplained and uncertain. Probably, we suggest, for a good reason. Humans make overriding discretionary choices and they blindly follow machines, often not knowing when either are taking place.[9] There is nothing about the HITL relationship that *ensures* quality decisions, or even makes them more likely. A human might trust a machine falsely just as they might override it falsely.

The AIA characterises regulating AI through a risk-based approach posing human arbiters as risk assessors and stop gaps to a varying landscape of risky implementations of AI without guaranteeing mechanisms that take into account the likelihood of the human fallibility of review.[10] This police and control approach precludes other approaches that marry the strengths of both actors in a decision making ecosystem. Iyad Rahwan and others have argued from a similar point of view for the need to establish the "... scientific study of intelligent machines, not as engineering artefacts, but as a class of actors with particular behavioural patterns and ecology."[11] Perhaps we are not quite at the intelligence parity point to see ADMs as full actors, but the imperative to approach the challenges ADM presents as a study of the ecology of the decision making environment would go a long way to developing a system of non-rivalling intelligences and empower the reciprocal relationships that trust, as we argue below, requires. We name this ecosystem: *cooperative intelligence*. Cooperative intelligence (CI) implies that ADMs can be combined in various hybrid constellations with manual (i.e. human) decision makers, taking into account alternatives to decisions suggested by the machine, or the reverse. *CI proposes developing the institutional framework to support an evidence-based appraisal of the relationships between human and machine intelligence to devise solutions that make administrative decision making aligned, efficient, and trustworthy*. A robust CI ensures that trustworthiness lives up to its name while safeguarding the agency and autonomy of citizens subject to these decisions. CI frames ADMs as actors within a larger behavioural ecology focusing on particular aspects of the relationship between humans and those systems. This is essential to understanding the mechanisms of a trustworthy decision space as supported by the literature in human computer interaction (HCI) and behavioural psychology literature, among others.[12] *How* humans make decisions, both alone and in concert with machines, matters.

While frameworks like the AIA are laudable for their ethical approach in *requiring* trust or trustworthiness, *how* this might be realised and designed should be embedded as part of the requirement. This article aims to deviate from a machine-as-threat conception of ADM and examines the concept of trust to reveal that it is a concept that relies on a *specific* relationship between humans, that we do not think can be (nor should be) strictly translated to the relationship between humans, or manual decision makers (MDMs) and their algorithmic counterparts. First, we define the space in which cooperation takes place, and the types of decisions that are involved in human-ADM interactions in an administrative setting. Next, we outline the conceptual parameters of trust, disambiguating trust with similar concepts and unpacking the kind of trust we think that the AIA and similar documents are after. Finally, we argue that trustworthy AI cannot be achieved by technology alone but requires a specific contextual setting that is characterised by an ecosystem of cooperative intelligence and we sketch a broad outline of how parallel processing tracks with built in blindness can ensure human-ADM cooperation.

14336/22 of 11. November 2022, interinstitutional file: 2021/0106 (COD.

6 See Crootof, Rebecca and Kaminski, Margot E. and Price II, William Nicholson, Humans in the Loop (March 25, 2022). 76 Vanderbilt Law Review 429 (2023), U of Colorado Law Legal Studies Research Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011, Available at SSRN: https://ssrn.com/abstract=4066781 or http://dx.doi.org/10.2139/ssrn.4066781 Other configurations bely this ambiguity see, Christina Wiethof and E Bittner, 'Hybrid Intelligence-Combining the Human in the Loop with the Computer in the Loop: A Systematic Literature Review', *Forty-Second International Conference on Information Systems, Austin* (2021).

7 AIA (n 5).

8 While the AI Act is intended to ensure trustworthy AI in general, the main content of the Act, apart from prohibiting certain specific uses of AI (see art. 5), most of the Act addresses so-called "high risk" AI. What constitutes high risk AI is identified through usage contexts (for example education, employment, essential social services) that are enumerated in appendices to the Act.

9 Marina Chugunova and Daniela Sele, 'We and It: An Interdisciplinary Review of the Experimental Evidence on Human-Machine Interaction' Center for law & economics working paper series (Social Science Research Network 2020) SSRN Scholarly Paper ID 3692293 <https://papers.ssrn.com/abstract=3692293>.

10 See for instance, 'Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment | Shaping Europe's Digital Future' (17 July 2020) <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment> accessed 3 March 2023.

11 Iyad Rahwan and others, 'Machine Behaviour' (2019) 568 *Nature* 477.

12 On human-machine interaction in the context of public administrative law, see especially, Jennifer Raso, 'Displacement as Regulation: New Regulatory Technologies and Front-Line Decision-Making in Ontario Works' (2017) 32 *Canadian Journal of Law & Society* 75; Jennifer Raso, 'Unity in the Eye of the Beholder? Reasons for Decision in Theory and Practice in the Ontario Works Program' (2019) 70 *University of Toronto Law Journal* 1; Leid Zejnilović and others, 'Algorithmic Long-Term Unemployment Risk Assessment in Use: Counselors' Perceptions and Use Practices' (2020) 1 *Global Perspectives*.

2. The problem space: the spectrums of decision-making and trust

In any administrative setting there are numerous arrangements of interactions and stages of decision making. These arrangements can be portrayed as a spectrum of interaction. On one side of the midpoint, lies manual decision-making process (MDM) and on the other, automated decision making. At the end points of each side lies the purest sense of their respective formats, pure MDM in which no digital apparatus is used at any point for the purposes of the decision itself and pure ADM, where complex decisions are made completely autonomously by robotic systems. Realistically in both cases, this would be rare to find. In MDM, decisions might still make use of digital organisational tools or databases, but if those systems use no form of recommendation or automation, they are true analogues of non-digital system. Likewise, for a decision to take place, even a fully automated system would need an input that originates from outside the fully automated system itself.[13] An administrative decision governed by law, whether manual or automated, is the result of an input in the form of factual information that is being processed in various ways. Most case handling involves some consequential human-machine interaction.

In-between these outer points there exists a potentially infinite amount of hybrid approaches. This would include, for example, automatic information retrieval, automatic processing of information, statistical visualisation systems, recommender systems, etc. We can use the example of a caseworker in a generic public administrative setting to flesh this out. As one travels from pure MDM towards pure ADM, a caseworker would use simple digital tools, e.g. such items as automatic spell check or simple automatic form filling from previous iterations of similar documents from the same user. Further along the spectrum is the introduction of digital templates with pre-configured text or perhaps a system that introduces information automatically from other electronic sources to the case worker. More fully automated decision-making processes might determine outcomes automatically with minimal to no input from the caseworker based on input provided by citizens via electronic systems, which citizens operate themselves (this might include a pre-application checklist with point scores for visas or state funding applications). Near the far end of the automation spectrum lie nudging systems, recommender systems, chatbots and similar designs based on machine learning from larger data sets providing a caseworker with discretionary choice of outcomes based on a score, multiple choice, or narrative recommendation of an algorithm.

In this varied space, decision making is best characterised by seeking to fit legal rules and principles, that almost always contain some discretionary elements, to a factual situation, that is often characterized by a social complexity that can be difficult to formally document in administrative procedure. Applying legal rules and principles to an individual citizen's case is therefore rarely straightforward.[14] It is by no means a new observation that as we hand over more work to algorithms, the "intelligence" used to perform a given task is often reduced to a mere ordering device, making decisions conform to certain parameters defined in advance. However, the difference between the artificial intelligence (as ordering device) and the human intelligence (as a truly context sensitive decision maker) is not as black-and-white as the machine-as-risk model presupposes. Whether run with or without computational machinery, bureaucratic institutions are created for the explicit purpose of handling large amounts of individual (but similar) decisions in an efficient and systematised manner via specialization and routinized decision-making processes.[15] In fact, the likening of bureaucracy to a machine stems from its likeness in scale and repetitiveness. It points to the specific, impersonal character associated with the bureaucracy: conformity to rules and formalities that prevails over personal relations and empathy.[16]

The extent of which this iron cage of rationality[17] is desirable and/or unavoidable in modern society can be debated. Our point here is that bureaucratic operations can (and in many cases, should) be translated into procedures that can be handled by machines in their respective capabilities. If a machine learning algorithm is trained on large datasets of public decisions in some domain of public administration and if the decisions are sufficiently homogenous, it seems likely that ADMs can be used to enhance efficiency and equality in public decision making practices.[18] The danger here, as often cited or alluded to, is that rule following becomes the placeholder of measured human discretion. Unlike a machine, a caseworker can react to the potential consequences of adverse events. However, the same charge has been levied in the reverse. There is continual cross-jurisdictional pushback to allow for and guard against human discretion in these processes.[19] With the discretionary good also lies

13 A decision on tax for example, relies on information about income, which must derive from a sum of money that is actually paid by an employer to an employee in return for labour. A decision to allow the construction of a new garage on someone's property must derive from the property owner deciding at some point to apply for permission to build a garage and submitting an application for permission. To grant a handicap friendly car to a person who has lost one leg presupposes that someone has made an application for such a car. Processing a speeding ticket can only happen if someone has actually been driving their car at too high speed. Only when there is an interaction between a provided input (factual information) and an output (a decision) that is not identical to the input, but is a result of processing and thereby of a transformation of the input to some output, does it make sense to talk about a decision. Hence even a fully automated decision system relies to some extent on non-automated events outside the system which at some stage must be put into the system.

14 See for instance on rule-breaking and interaction with perceived red-tape, Randall S Davis and Stephanie A Pink-Harper, 'Connecting Knowledge of Rule-Breaking and Perceived Red Tape: How Behavioral Attribution Influences Red Tape Perceptions' (2016) 40 *Public Performance & Management Review* 181.

15 The effects of bureaucratic organizations in terms of streamlining the procedures of work is well known. In the most emblematic study of bureaucracy, Max Weber's essay of the same name, Weber even compares the bureaucracy to that of a machine: "The fully developed bureaucratic apparatus compares with other organisations exactly as does the machine with the non-mechanical modes of production" Max Weber, *Economy and Society: An Outline of Interpretive Sociology* (Bedminster Press 1968) p.973 . See also a new introduction to Weber's text in: Tony Waters and Dagmar Waters, 'Bureaucracy' in Tony Waters and Dagmar Waters (eds), *Weber's Rationalism and Modern Society: New Translations on Politics, Bureaucracy, and Social Stratification* (Palgrave Macmillan US 2015) <https://doi.org/10.1057/9781137365866_6>.

16 Elin Wihlborg, Hannu Larsson and Karin Hedstrom, '"The Computer Says No!" -- A Case Study on Automated Decision-Making in Public Authorities', *2016 49th Hawaii International Conference on System Sciences (HICSS)* (IEEE 2016) <http://ieeexplore.ieee.org/document/7427547/>; Sofia Hina Fernandes Da Silva Ranchordas, 'Empathy in the Digital Administrative State' (2022) 71 DUKE LAW JOURNAL 1340.

17 See Weber, (n. 15).

18 In a report from 2018 issued by the Danish Moderniseringsstyrelsen (a department under the Danish Ministry of Finance), text analytics and predictive modelling is estimated to have a large potential for repetitive and text-intensive case handling. See: <https://modst.dk/media/29941/rapport-om-kortlaegning-af-analytics-i-staten.pdf> at p. 44.

19 Joana Mendes, 'Discretion, Care and Public Interests in the EU Administration: Probing the Limits of Law' (2016) 53 *Common Market Law Review*.

the discretionary evil. Retaining a human in the loop does not guarantee a decision-making process is free of strict rule following. It may merely obscure it through the veil of an assumed, active human discretion even when it is plainly not the case. The appearance shouldn't be enough. Given the enormity of demands on decision maker's time, the scale of information used to make coherent decisions, various organisational pressures on conformity, among other demands, we are not convinced that this assumption is realistic or warranted.

This element in decision making warrants restatement: just because a decision is made without the involvement of machines, it does not guarantee that the decision is the product of independent, fair, or unbiased judgment on the part of the decision maker. Just as there is a continuum between the manual and the automated, so too is there a continuum between rule and discretion:

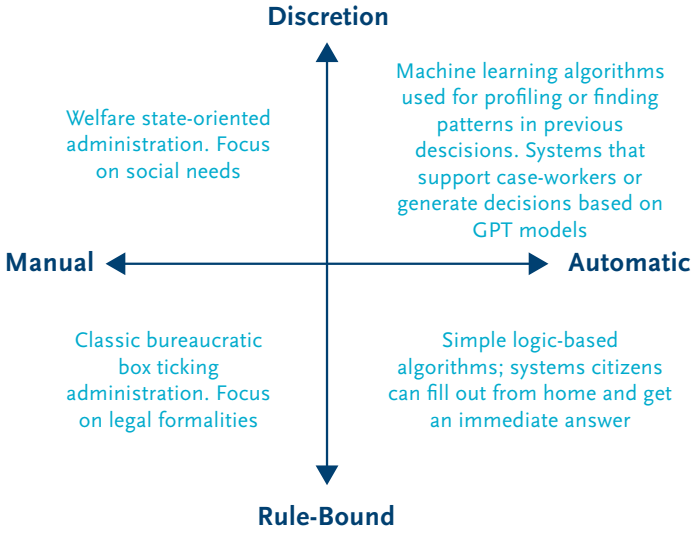

Figure 1. illustrates the two dimensional space where the horizontal axis shows the Manual-to-Automatic continuum and the Vertical axis the Rule-Bound-to-Discretion continuum.

A fully manual decision may be fully rule-governed and in that sense no different from an automated decision. A factory worker who pulls a lever every time a bell rings is performing a manual job, but does so in a way that leaves no room for discretion - except perhaps in the very minor sense of choosing the speed at which the lever is pulled and how fast after the bell rings the pulling is executed, or not pulling it and getting fired which in all rights is never much of a choice. Similarly, a case worker may grant or refuse to grant some permission, service or benefit to a citizen in a quasi-automatic way by checking whether the conditions of some specific rule is fulfilled in the case before them, going through a checklist or a point system. In such situations, even if the decision making is fully manual, bureaucratic routine can have the same effect on the decision-making procedure as a set format imposed by a computerized system[20]. This is to say nothing of the rule following of unconscious bias, expectation, perceived desert, personal grudges or the like.

What allows discretionary systems to function in spite of these limitations (and what ethicists and lawmakers are demanding out of ADM) is the purgative power of trust. How can a person or society at large trust that discretionary decision making follows the demands of equality while also being sensitive to personal context, need, or exception? And who is it that we are trusting? The decision maker, the institution, the rule makers? The answer to these questions mean something for ADMs to be deemed trustworthy and it is not solved by human discretion. The problems exist prior to and after the introduction of AI into administration.

The notion of trust encompasses the acceptability and correctness of the ultimate decision of the trustee (either in its personal or metonymic sense) and the trustor's attitude towards it. Every trust situation retains a certain degree of risk and vulnerability - on the part of the trustor. What is needed here is a proxy for this relationship between trustee and trustor that makes sense in the context of that decision making in a public administrative setting. Whether it is a fully manual, fully automated, or any variation on the spectrum, the acceptance of "the system" (the administrative bureaucracy) will hinge mainly on an individual's and their society's ability to trust that the decision outcomes are both correct (with regard to their fidelity to reality and to their legality) and remediable. However, there is an important corollary to point out here.

Even if we were to have full trust in a decision maker and the system in which they were embedded we would not have a system that was devoid of error, misrepresentation, or bias. There are many occasions where distrust in an ADMs is precisely what is needed for a larger ecosystem of trust. We will return to this notion below distinguishing between trust and reliability, but we can use a short, very recent example to illustrate the point. Open AI's recent language model ChatGPT has caused a firestorm in predictive doomsaying. Of the many industries it is said that will be wholly upended by the ability to replicate human like text creation is that of the university essay. This is a good analogue for the trust scenario we are discussing here. Much like the rule applying caseworker, for university law students[21] writing an essay for an exam can involve applying a rule to a given scenario given the rules, principles and caselaw learned in class. A system like ChatGPT produces human like output while not necessarily being entirely accurate and definitely does not allow the student to be graded on a measure of their own knowledge. This has led to the inevitable clickbait headlines that it will render the university essay useless. Whether or not this is true, this scenario can teach us much about the function of trust within a hybrid system.

Let's imagine a scenario where a student has enlisted a chatbot to produce her essay and believes that the output is correct and turns it in. The obvious downside for the student is that the bot has output grammatically correct nonsense, or authoritative but trivial and meaningless text Trust here, functioned as a foil to an accurate outcome. However, if the student found the system untrustworthy, or more accurately unreliable, to deliver correct answer, their time and effort can instead be spent on fact checking and making sure that the answers are context sensitive and nuanced. In fact, this behaviour is incentivized. It would enable a cooperative effort between student and chatbot that would deliver the kind of behaviour that the insti-

20 The infamous and brutal example of this is famously described by Hanna Arendt in Hannah Arendt and Jens Kroh, *Eichmann in Jerusalem* (Viking Press New York 1964).

21 For the sake of the analogy, we are referring to law students or social science students generally. The analogy of rule application would make much less sense for say a creative writing or literature student.

tution is seeking to deliver: a mindful human who goes beyond rote rule application and critically engages with the material as produced by the machine. A win-win. We believe this analogises well to the case worker working with ADMs. Trust ecosystems can develop and encourage certain desirable behaviours and discourage others. Given the spectrum between the poles on trust, reliability and other similar concepts, we must ask: when is trust practical to a decision and in which instances should be fostered? Or, when is a decision "trustworthy"? To answer this, we must first unpack the concept.

3. The conceptual parameters of trust

Trust is a mental "attitude"[22] amounting to a "bet about the future contingent actions of others."[23]The notion of trust has a specific meaning, but it is related to it a family of terms that includes "knowledge" and "belief". Like for "knowledge" and "belief", we do not *choose* to trust. Once we hold a relevant belief, that belief informs our degree of trust.[24] Trust has also been positioned as a phenomenon that challenges the usual division of mental phenomena into cognitive (reason), affective (emotional), and behavioural (volitional choice) categories, perhaps belonging to all of these. "[Trust] is based on what we believe about another person or agency, it requires that we feel safe in their hands, and it usually involves a voluntary act of entrusting, or placing trust."[25] Although essentially a cognitive phenomenon, trust crosses the usual distinctions. As Weigert and Lewis have argued, these three components of trust are all merged into a unitary social experience, but the relative importance of each provide grounds for differentiating subtypes of trust.[26] In this conception, trust can neither be fully understood without all these components nor reduced to only one of them; for example legislation like the AIA that reduces trust to the manifestation of trusting behaviour. Moreover, categorising trust without one of these aspects may make such categorisation conceptually deficient, as it might not amount to a concept of trust, but rather of neighbouring ideas such as confidence, reliance, faith, or gullibility. Since trust as a phenomenon always requires volitional choice, or in other terms discretion or consideration, talking about it purely in relation to AI, machine learning or ADM (eg. as "trustworthy AI") without any human contextualisation, is a misconception. When a purely automatic output is given, there is no space to talk about trust of that output. Instead, reliability would be a correct term for what we observe and experience.

Reliability is often used to describe and even define trust.[27] Katherine Hawley contends that trust involves reliability in that a common trait of a trusting relationship is practical reliance on the trustee.[28] However, mere reliance is distinguished by existing as behaviour. Trust, in this view, occurs prior to the decision to rely on[29] and the act of reliance itself – *trust is not an action, but an attitude that comes before and explains an action*. Besides being a part of the structural composition of trust, relying does not always require the prior attitude of trust. Thus, there is a more fundamental differentiation between them besides the temporal antecedence of trust. Trust is also distinct in a moral sense. Trust has a moral quality that is not necessarily present in reliance.[30] Reliance may be a purely pragmatic way of relating to others (or other things).[31] Both trusting and relying on someone/something involves presupposing that they will act as expected. Trust involves such a reliance plus some additional factor.[32] We also rely on things (i.e. a clock) and other inanimate objects, but trust is characterised by the attribution of responsibility and agency in upholding the trusting relationship. So talk of trustworthy machines in this sense is a definitional non-starter. Since machines are devoid of moral character and responsibility, we can only relate to these by either relying or not relying on them.[33] Trustworthiness on the other hand can only be a character of a decision ecosystem, which involves human agency and hence responsibility.

Richard Holton has described this difference between trust and reliance in that trust involves taking a participant stance towards the person you are trusting. When you trust someone to *do* something, you rely on them to do it, and you regard that reliance in a certain way – you have a readiness to feel betrayal should it be disappointed, and gratitude should it be upheld.[34] This is a part of treating them as a person[35] – giving them agency and discretion and attributing responsibilities, but accordingly also possible blame or merit. Note that blame, betrayal, gratitude, and merit are again the manifestations of the affective and moral components of trusting; they are strong emotions after one feels they are let down, often in contexts involving moral and value-laden normative expectations.

This additional factor in differentiating trusting from mechanical reliance has been described by Hawley as having something to do with heightened expectations in trusting and your reaction if the trustee lets you down.[36] She proposes that this heightened expectation comes from a commitment of the trustee in what she conceptualises as the commitment account of trust: to trust someone to do something is to believe that she has a commitment to doing it, and to rely upon her to meet that commitment.[37] These commitments are best captured in the form of promises, but not only. Commitments can be explicit or implicit, weighty or trivial, conferred by roles and external circumstances, default or acquired, welcome or unwelcome.[38] Commitments trigger moral responsibility when they are not met and reactive attitudes on the part of the trustor – notably resentment, feelings of betrayal, and blame. This also explains why we can rely on inanimate

22 Karen Jones, 'Trust as an Affective Attitude' (1996) 107 Ethics 4; Niklas Luhmann, *Trust and Power: Two Works* (Wiley 1979) 27.
23 Piotr Sztompka, *Trust: A Sociological Theory.* (Cambridge University Press 1999) 25.
24 Russell Hardin, *Trust* (Polity 2006) 17.
25 Annette Baier, ´What is Trust?` in David Archard and others (eds), *Reading Onora O'Neill* (Routledge 2013) 177; Annette Baier, *Moral Prejudices: Essays on Ethics* (Harvard University Press 1994) 132.
26 J David Lewis and Andrew Weigert, 'Trust as a Social Reality' (1985) 63 *Social Forces* 20, 969–970.
27 See for example: Annette Baier, 'Trust and Antitrust' (1986) 96 *Ethics* 231, 234; JM Barbalet, 'Social Emotions: Confidence, Trust and Loyalty' (1996) 16 *International Journal of Sociology and Social Policy* 75, 77.
28 Katherine Hawley, *How To Be Trustworthy* (1st edn, Oxford University Press 2019) 2.
29 Larry E Ribstein, 'Law v. Trust' (2001) 81 Boston University Law Review 553, 560.
30 Olli Lagerspetz, *Trust, Ethics and Human Reason* (Bloomsbury Academic 2015) 15.
31 Cynthia Townley and Jay L Garfield, 'Public Trust' in Pekka Mäkelä and Cynthia Townley (eds), *Trust: analytic and applied perspectives* (Brill 2013) 97.
32 Hawley (n. 28) 8.
33 See also: Mark Coeckelbergh, 'Can We Trust Robots?' (2012) 14 *Ethics and Information Technology* 53; Mark Ryan, 'In AI We Trust: Ethics, Artificial Intelligence, and Reliability' (2020) 26 *Science and Engineering Ethics* 2749.
34 Richard Holton, 'Deciding to Trust, Coming to Believe' (1994) 72 *Australasian Journal of Philosophy* 63, 67.
35 Holton (n 34) 67.
36 Katherine Hawley, *Trust: A Very Short Introduction* (Oxford University Press 2012) 5.
37 Hawley (n 28) 9.
38 Hawley (n 28) 10.

objects, but not trust them – they do not possess human agency that is capable of assuming responsibility or be responsive to such normative expectations. This doesn't stop us, however, from anthropomorphising inanimate objects in order to speak about (or even demand) trust in relation to our interactions with them. Reliance, besides being the manifestation of the behavioural component of trust and without a moral, commitment-triggered connotation can also be placed on inanimate objects. From this point of view, this added commitment differentiates trust from reliance, and it triggers responsibility, liability and accountability if not met.

The differing ecosystems where this distinction could be applied would change depending on the actors and the relationships we are trying to bring "trustworthiness" to. Sociologist Diego Gambetta has defined trust as a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to be able to monitor it) and in the context in which it affects his own action.[39] When we say we trust someone or that someone is trustworthy, we implicitly mean that the probability that he will perform an action that is beneficial or at least not detrimental to us is high enough for us to consider engaging in some form of cooperation with him.[40]

As seen from these definitions, trust is never an abstract or non-conditional notion. To unpack trustworthiness one must ask: "trust between whom and in what?". It is a three-place relationship, where "trustor trusts trustee in doing/with respect to something"[41], where this thing is the content of the relation – the entity in which trust is directed.[42] In the context of ADM then, it is necessary to clarify the agents (subjects) in the trust relationship. In the bureaucratic caseworker example, the trust relationship model might be seen as this: citizen A trusts caseworker B in doing its task X, where X might be the act of deciding on a benefit, resolving an asylum application, granting or refusing a permit, etc. But the trust relationship might also be wider, where citizen A trusts institution B in handling all the cases relating to task X. In this case A does not trust a specific person, but a group of (anonymous) persons metonymically associated with the institution.

Trust also exists *within* institutions, for example, as a trust relationship between colleagues A and B. In this case X would be, for example, the act of exchanging advice and support in how to deal with a case. Trust could also be about caseworkers and their manager(s). In the context where a caseworker may trust or distrust an AI/ADM system, the situation becomes even more complex as trust in this case is a multidimensional issue. Person A in this case may be a citizen, a caseworker or a manager with responsibility for AI/ADM and B may be a caseworker, an institution, a manager and/or a provider of AI/ADM, where X will be the acts performed wholly or partly by the AI/ADM system, but in the name and responsibility of B. Whether AI/ADM is perceived as trustworthy depends very much on the quality of the human relationships within which AI/ADM will be used, and how they are perceived.

Actors able to manifest and receive trust need not always be people, although this is often the case. What is important is the notion of human agency, as trust is a mental phenomenon – trust can be traced back to some form of human cognitive conditions, and thus it needs to relate back to actors, to whom it is possible to ascribe expectations and actions meaningfully.[43] Collective and non-human entities can still be on the receiving end of trust (the trustees or the objects). The literature uniformly agrees on the existence of what is labelled with the varying names of abstract[44], system(ic)[45] or institutional trust.[46] These are, in essence, trust being directed at the adequacy, honesty, and competence of systems and institutions to live up to their commitments and carry out their duties rather than requiring a cognitive agent.[47] Trust and institutions are inherently intertwined. Institutions can be seen as bases, carriers and objects of trust. Trust between actors can be based on institutions, trust can be institutionalized, and institutions themselves can only be effective if they are trusted.[48] It is important to be analytically clear what the target of trust is in an institutional setting: is it the institutional arrangement itself or another actor's behavior which is constitutively shaped by these institutional arrangements, although in reality it is often difficult to make this distinction.[49] What is important here, is that institutions are both a source and an object of trust.[50] Institutional-based trust is tied to formal societal structures which generalize beyond a given transaction, and beyond specific sets of exchange partners and becomes part of the external world known in common.[51]

39 Diego Gambetta (ed), *Trust: Making and Breaking Cooperative Relations* (B Blackwell 1988) 217.

40 Gambetta (n 39) 217. However, there are still more elements to be considered. A much-cited study on interdisciplinary understandings of trust has proposed an integrative definition: "Trust is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control the other part." This willingness to be vulnerable has sometimes also been referred to as a willingness to depend on another actor with regard to something.

41 Stephen Alexander Rompf, *Trust and Rationality: An Integrative Framework for Trust Research* (Springer 2015), 32; Russell Hardin, *Trust* (Polity 2006), 19; Eric M Uslaner (ed), *The Oxford Handbook of Social and Political Trust* (Oxford University Press 2018) 6; Russell Hardin, *Trust and Trustworthiness* (Russell Sage Foundation 2002), 9; Katherine Hawley, *How To Be Trustworthy* (1st edn, Oxford University Press 2019) 25.

42 Rompf (n.41), 32.

43 Guido Möllering, *Trust: Reason, Routine, Reflexivity* (First edition, Elsevier 2006) 7.

44 Barbara A Misztal, *Trust in Modern Societies: The Search for the Bases of Social Order* (Polity Press; Blackwell 1996) 72.

45 Luhmann (n 22) 50; Misztal (n 44) 133.

46 Möllering (n 43) 71.

47 Hawley (n 36) 99.

48 Möllering (n 43) 74.

49 Reinhard Bachmann, 'Institutions and Trust' in Rosalind Searle, Ann-Marie Ingrid Nienaber and Sim B Sitkin (eds), *The Routledge companion to trust* (Routledge 2017) 219.

50 Lars Fuglsang and Søren Jagd, 'Making Sense of Institutional Trust in Organizations: Bridging Institutional Context and Trust' (2015) 22 *Organization* 23, 27.

51 Fuglsang and Jagd (n 50) 26. Lynne G. Zucker has distinguished three different sources of trust: process-based trust, characteristic-based trust and institutional-based trust. Process-based trust is based on concrete experience from past exchange that may be either first-hand or third party experience passed on to the trustor by 'trust intermediaries'. Process-based trust may cumulate into reputation that may enter into the trust decision process performed by the trustor. Characteristic-based trust is independent of a concrete exchange experience; the sources of this kind of trust are personal characteristics such as age, sex or belonging to a particular ethnic community. Institutional-based trust is considered as being generated more diffusely in a wider network of relationships. Sources of institutional-based trust may be traditions, professions, certifications, licenses, brand names or membership in associations. See, Lyanne G Zucker, 'Production of Trust: Institutional Sources of Economic Structure, 1840-1920' (1986) 8 *Research in Organizational Behavior* 53.

Trust in an institution means confidence in the institution's reliable functioning, but this has to be based mainly on trust in visible controls or representative performances rather than on the internal workings of the institution as a whole.[52] Institutions facilitate trust by their inbuilt rules, roles and routines – the bases of trust in an institutional setting.[53] Where institutional trust exists, the trustor and the trustee refer to institutional safeguards in their decisions and actions and can thus develop trust without having any prior personal experience in dealing with one another.[54] In other words, institutions shape and channel actors' expectations so as to allow for trust-building in many situations where otherwise there would be no trust.[55] Outside of the general prescriptions for data representativeness, correctness, etc., instruction on human oversight and transparency as well as logging requirements etc. there is little concrete clarification in the AIA, or the ethical guidelines on AI[56], which human-computer design should be used to retain trustworthiness or how it should be measured or understood.

This section has clarified why The EU's attempt to further trustworthy AI through the AIA and its underlying ethics guidelines not sufficient on its own to ensure that public administration remains or becomes trustworthy when AI is introduced to these institutions. Even so, we might imagine that the most tractable way to overcome the multitude of relationships that could exist to develop trustworthy AI worth its name is to rely on an institutional point of view where, trustor A (a citizen) trusts in B (a public institution) whose task it is to perform a decision relevant to A. Trust can be had in an institution, which performs an evaluation of eligibility for public assistance or similar, whether or not that institution uses AI to perform or assist in the performance of evaluation. It is the institution within which the technology is embedded – not the technology itself – that is trusted. However, it remains true that the reliability of an AI system used by an institution to perform its functions, may affect the level of trust attributed to the institution. If the institution acts in a way that comes across as irresponsible, for example by introducing an AI system that automates decision-making and gives rise to a sudden and significant increase in flawed decisions (or is even perceived as doing so), then obviously the trust in that institution is undermined.[57] Once this need for contextualization is understood, it becomes clear that rather than "trustworthy AI" we should talk about "trustworthy use of AI". Given this, it is questionable whether the AIA is indeed suited for promoting trustworthy (use of) AI. For example, Article 14(1) on human oversight, requires that:

> High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.[58]

This is fundamentally a requirement aimed at AI developers (but with responsibility for "providers") that the AI system must be set up in such a way that it is possible to carry out a review of its operations. This we think is quite uncontroversial, and it neither adds nor subtracts from the trust a citizen may have in an institution. The existence of an "appropriate human-machine interface tool" is no guarantee that humans will in fact oversee the operation of the AI system in question or do so in a way that engenders trust in the institution. This is not remedied by 14(2)[59] which is essentially an instruction to those who construct the human-machine interface tools that oversight mechanisms must be aimed at those automated functions in the system that may lead to harm for humans.[60] This is not surprising given that the AIA has the character of product regulation,[61] and whether citizens trust institutions or companies who use AI as part of their operations will depend on how users engage with, understand or are fed with information about these AI systems. The very existence of an oversight window, does not in and of itself provide for trust in AI.

The closest we come to a regulation of how a human in the loop is supposed to function in practice is in Art. 14(4). Art. 14(4) sets out five functionalities that must be built into the oversight window. These five functions are described in terms of the ability of human oversight to perform certain cognitive and manual tasks.[62] There is

52 Guido Möllering, 'Trust, Institutions, Agency: Towards a Neoinstitutional Theory of Trust' in Bachmann Reinhard and Zaheer Akbar (eds), *Handbook of trust research* (Edward Elgar 2006) 365.

53 Möllering (n 43); Möllering (n 52) 362.

54 Reinhard Bachmann and Andrew C Inkpen, 'Understanding Institutional-Based Trust Building Processes in Inter-Organizational Relationships' (2011) 32 *Organization Studies* 281, 282.

55 Bachmann and Inkpen (n 54) 297. It should be borne in mind that trust, besides being fragile (i.e. it is easier to break-off trust than to build it), also has easy spill-over and transformational effects. That means that trust or distrust between individual actors (i.e. between a citizen and a caseworker) can easily transform into trust or distrust in the institution, which the individual trustee represents. Sociological trust barometers often inquire into aggregate citizens trust in different state institutions, which is also based on individual experiences and hearsay. For this reason trustworthiness by design, while laudable, is a short-sided and anaemic approach to the relationships implied by digital administration.

56 See <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.

57 Sonja Bekker, 'Fundamental Rights in Digital Welfare States: The Case of SyRI in the Netherlands' [2021] Netherlands Yearbook of International Law 2019: Yearbooks in International Law: History, Function and Future 289.

58 AIA (n.5).

59 As a central part of the AIA's regulation of high risk AI, art. 14 plays a key role in the attempt to satisfy the demand for trustworthy AI. This comes out clearly in the explanatory memorandum that accompanies the AIA: "This proposal aims to implement the second objective for the development of an ecosystem of trust by proposing a legal framework for trustworthy AI. The proposal is based on EU values and fundamental rights and aims to give people and other users the confidence to embrace AI-based solutions, while encouraging businesses to develop them. AI should be a tool for people and be a force for good in society with the ultimate aim of increasing human well-being. Rules for AI available in the Union market or otherwise affecting people in the Union should therefore be human centric, so that people can trust that the technology is used in a way that is safe and compliant with the law, including the respect of fundamental rights." (European Commission proposal for AIA (COM(2021) 206 final; 2021/0106 (COD)), 21 april 2021, p. 2).

60 "human oversight shall aim at preventing or minimizing the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter" ibid.

61 Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act' [2021] arXiv:2107.03721 [cs] <http://arxiv.org/abs/2107.03721>.

62 To "fully understand the capacities and limitations of the high-risk AI system and ... [to make the human] able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible"; to "remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons"; to "be able to correctly interpret the high-risk AI system's output, taking into account in particular the characteristics of the system and the interpretation tools

no doubt that the intention behind this is to evoke a sense of trust through the association with "human control". Humans understand other humans better than they understand machines and ensuring human control over machines will undoubtedly be seen by many as an important safeguard against potentially inhuman operations by machines. AI systems then, must be designed in such a way that humans can understand the machine, and thereby be able to disagree and ultimately say override the machine. When to say yes, when the computer says no? One could imagine any number of scenarios in which regulators would want to incentivize this kind of discretion to ameliorate instances of data bias or feedback loops. It would guard against the tyranny of blind rule following and add flexibility into the system. However, the insertion of a human in the loop, if wholly untrusting of the machine, defeats the purpose of automation in the first place. What is the point of operating a machine system if every time it outputs an answer, there is a discretionary speed bump in the way? For a balance to be struck, we must design a model for balancing these competing requirements during the oversight window to generate the appropriate amount of cooperation.

4. Institutional Trust via cooperative intelligence

Given the variability across different types of tasks, scenarios, contexts and actors, whilst designing for institutional trust, an ADM hybrid model would need to contain both an element that would amplify trust in system recommendations/decisions made by machines on their own alongside an element that would incentivise distrust and intervention by human arbiters. Such a system would need to be both *outcome centric* and *human centric*. By outcome centric, we mean that what matters is functional correctness and access to human remediesif the AI system does not meet the functional correctness requirement. What matters is not the medium of the decision maker (AI or human) in itself, but the function of the system as a whole. By human centric, we mean that such systems must retain the flexibility of a purely human system, so that it can adapt to unknown situations and regulatory change. Such a system would make use of the strengths of both types of participants (AI and human) and would optimise its function in light of the realities of how they interact with each other in various situations/scenarios.

Safeguards that are simply based on a human stopgap understanding of ADMs are not enough. While they may provide a feeling of retained authority they also put humans in a position of overreliance on automated decisions in the form of "automation bias" or "automation-induced complacency" which is "the failure to appropriately monitor [and intervene with] automated support."[63] This situation becomes more problematic in a hybrid system where it is not exactly clear who has ultimate responsibility for the decisions.[64] In a discretionary system, someone must be held responsible for those decisions and be able to give reasons for them. There is a legitimate fear that a black box system creates a scenario that lacks responsibility, even when used in coordination with a human counterpart or oversight. This is the fear of the rubber stamp: even if a human is in the loop, the deference given to the machine is so much that it creates a vacancy of accountability for the decision.[65] While compelling - in the sense that reasons and the responsibility for those reasons are highly linked to our sense of legality and fairness - one can ask, what has actually changed in terms of reasons and responsibility by including the machine?

Imagine a fully manual system where a caseworker is deciding on an application for unemployment benefits. They take in the relevant data, follow their training, perhaps ask advice from a superior or co-worker, weigh options, finally arrive at a decision, and deliver the outcome with its (usually cursory and templated) reasoning to the applicant citizen. The public service who hired this civil servant is ultimately responsible for the decision, not the civil servant themselves. The institution had the responsibility of vetting and hiring the decision maker, and usually, that decision maker is barred from any personal liability if it is found they have followed the rules prescribed for making a decision. The legal culpability of the public body in any challenge or remedial process will ultimately rely on the legislation regarding the reasoning for that type of decision and an assessment of the adequacy of the safeguards in place for a particular type of harm. With all of this in mind it is hard to see a HITL system that would be able to balance these barriers against each other to provide context-specific yet equally applicable rules. How would they know when the computer is wrong? While there may be jurisdictions out there that would accept "I'll know it when I see it"[66] type reasoning, most would not allow this as an adequate safeguard. One way around this dilemma is to look at how humans do in fact interrogate this question.

The need for more studies on the intricacies of these types of hybrid interactions cannot be understated. In a recent review of experimental evidence of human-computer interaction it was found that during interactions in decision making scenarios between humans and automated agents, the context of the task or differing role of the participants matter considerably and can lead to opposite results of relying on automated decisions.[67] For instance, there is a dampening effect of emotional and social response to decisions made by ADMs (positives are less positive negatives are less negative) while at the same time they were "willing to engage with automated agents in contexts perceived as analytical or objective, [but] they seem reluctant to do so in more social or moral contexts."[68] Furthermore, they were still more comfortable revealing personal information to a machine as compared to a human due to a lack of concerns of social image repercussions..[69] Similarly, the reviewed studies showed context specificity in delegating tasks to ADMs, particularly in terms of the distribution of agency between human and machine where delegating authority and responsibility differently led to either aversion to the system or overreliance on it, and that "the empirical evidence seems to suggest that humans are averse to fully giving up their decision authority, yet appreciative of automated advice when they retain - or when they *feel* that they retain - the ultimate authority over the decision."[70]

and methods available"; to "be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system"; to "be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure".

63 Chugunova and Sele (n 9).

64 A related, but more legal technical problem in regards to the introduction of AI in public administration is the question of when exactly a decision is made. Associated to this is also the problem of delegation. If a private IT developer designs a decision-system for a specific group of public decisions, does this mean that those decisions have been delegated from the public administration to the IT developer? We shall not pursue these questions in this paper.

65 Wihlborg, Larsson and Hedstrom (n 16).

66 See Justice Potter's famous quip, Jacobellis v. Ohio, 378 U.S.S.C. 187 (1964).

67 Chugunova and Sele (n 9).

68 Ibid.

69 Ibid.

70 Ibid, *emphasis added*. Continuing with the caveat that, "However, this

In the interim, we suggest that one way to ensure this kind of set-up is to blind the human co-worker of ADMs to knowledge of the compositional status of their counterpart; that is, if they are working with a human or a machine. Mirroring the ethos of its blindness, we suggest designing this interaction along the lines of a quasi-Turing test.[71] We are not aiming to answer the question whether machines can think or even whether machines can make decisions. Instead, we are interested in a way of testing the extent to which administrative decision making can be partly automatized by algorithmic programs while avoiding automation bias or automation fear and ultimately promoting the kind of institutional trust envisioned by proposals such as the AIA.

Following the Turing model, an administrative body could implement an experiment with algorithmic decision making in a way that would make use of a blind assessor. This could be done in the following way: A certain percentage of the entire case load could be given both to a human caseworker and to an algorithm. Both the human caseworker and the algorithm would produce a decision draft for the same case. Both drafts would be sent to a human evaluator (i.e. an official, who finalizes and signs off on the decision). Furthermore, formats for issuing drafts could be formalized to reduce the possibility of guessing, merely by recognizing the style of the drafter's language, who is who. The advent of advanced generative language models like the aforementioned ChatGPT makes this step more feasible without too much work on formalization. In this set-up, the human evaluator would not know which draft came from the algorithm and which came from the caseworker. This would leave it for the human evaluator to choose which decision draft was the most convincing as a candidate for a decision and then endorse this as the final decision in the case at hand. With such a set-up it would be possible to test out the real functional performance of an AI system and simultaneously ensure that the system preserves a continuous quality check by being measured against existing standards of human performance. While we explain this in a simple human versus machine quality test there is no reason to limit the arrangement to be perfectly equal in the amount of cases allocated on either side.

The overall idea is to use ADM in a set-up with human interaction, thereby creating a hybrid model for administrative decision-making that relies on ADM for scale and manual control for supervising and adjusting data input and output of the ADM to counter bad data and feedback loops. The model can be applied, we propose, to any area of public administration where there are a large pool of previous decisions that are generally considered to be correct[72] and which are available in a format that allows for these decisions to be fed to a machine learning algorithm for the purposes of training ADMs to make decisions in new cases of the same kind. Once an algorithm has picked up the relevant relations between input (facts and law) and output (decision), the next step will be to build a model which uses the learned relationship between input and output to predict the outcome of new cases, given a specific constellation of facts and law.

Previous research on case law prediction indicates that it is possible to realize correct prediction of precedent implemented as a recommender system even across a broad field of legal areas.[73] We propose a similar system but implemented as a recommendation system for decisions of new cases within a narrow and dense field of legal regulation (i.e. a specialized area of law where many decisions are made frequently). Using common bureaucratic practices, where decisions are built on pre-existing templates, ADMs constructed as a recommender system could produce a decision proposal in the template by using data in the form of pre-existing decision practice. It could be set up to show one or more similar decisions from the decision database on which the ADMs operate. A human caseworker could then easily check whether the new decision would be qualified as satisfactory and could then finalize the decision by approving it.

We propose to split the overall decision making process in *three tracks*. Track one will be almost fully automated. Track two will be fully manual. In track three, the same cases will be given to both ADMs and manual caseworkers in parallel. Then a "blind" human evaluator at the end of track three will pick what they consider to be the best/correct decision. The entire decision ecosystem carries the tracks over four phases:

Phase one: Calibration

Track one and two are made identical in terms of form. This includes harmonizing the decision format and the language form used in decision drafts. A decision is always based on some application, request, observation, report or some other data source. It is crucial that the quality of data is checked and most often this requires manual case work.[74] At the case preparation phase where the data input is qualified, automation should not be introduced, unless data is drawn from reliable pre-existing databases, and there is no need to query or add to these data.

Phase two: Allowing Drafts

This phase is about testing the ADMs functional quality. We suggest that a recommender system could be used in the trial period to perhaps produce more than one ADMs output for each case. The system

hypothesis so far rests on a small number of studies and some corollary results, while studies which purposefully investigate the impact of this factor remain lacking... Without such studies, researchers run the risk of focusing only on one kind of behavior, and extrapolating it to too many situations,".

71 Alan Turing, in a paper written in 1950 , sought to identify a test for artificial intelligence. In the paper, Turing asked the question: "Can machines think?" and the test he devised for answering this question consisted of a set up in which (roughly explained) two computers were installed in separate rooms. One computer was operated by a person – the other was operated by an artificial intelligence system (a machine). In a third room, a human evaluator was sitting with a third computer. The evaluator would type questions on his computer and the questions would then be sent to both the human and the AI in the two other rooms for them to read. They would then in turn write replies and send those back to the evaluator. If the evaluator could not identify which answers came from the person and which came from the AI (the machine), then the AI would be said to have shown ability to think.

72 It should be noted that there is no way to systematically check for correctness since there is no objective standard against which decisions can be measured. It is worth observing though that decisions that was considered correct at the time it was issued could be considered wrong at later point in time. There are two primary reasons why decisions could be considered legally incorrect: 1) legislative change could have been introduced or 2) an administrative practice could have been overturned as legally invalid by a court of law. In both cases it should be considered to remove such decisions from the training set.

73 See Enys Mones and others, 'Emergence of Network Effects and Predictability in the Judicial System' (2021) 11 Scientific reports 1.

74 Data availability, data quality etc. is also very important when considering whether a decision-making process is at all suitable for automation.

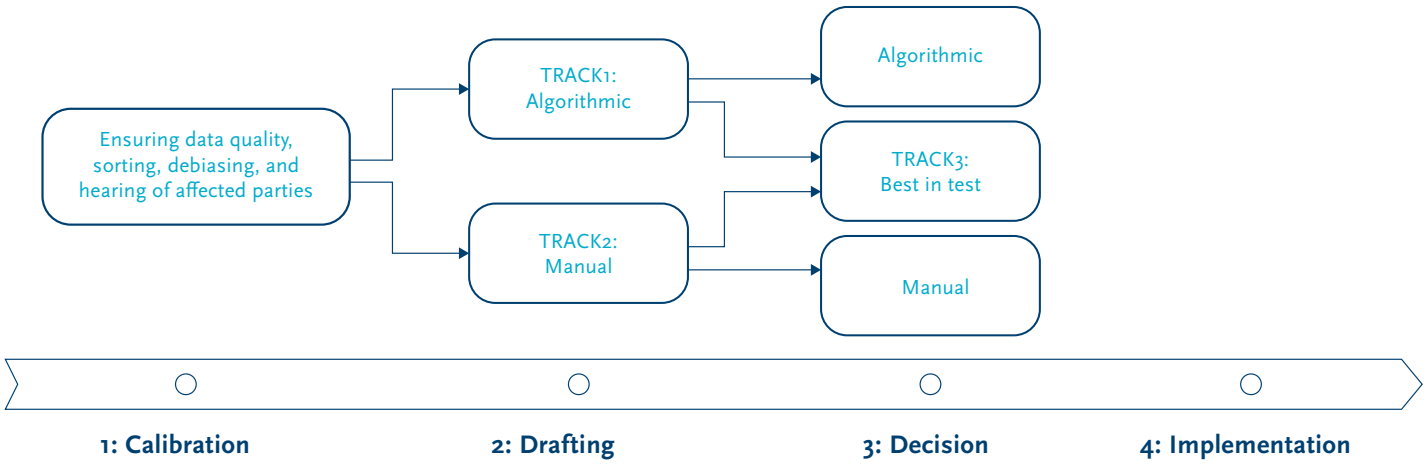

Figure 2. illustrates how an algorithmic support system that aims at meeting human standards for the writing of decision documents can be gradually phased into an organisation responsible for decision-making.

could then learn further by being informed about which of several recommended decisions are being chosen as the better decision. ADMs are still not put to use for actual decision making in this second step. Only the manually handled cases are used for actual decision making.

Phase three: Allowing Decisions

After the learning period in phase two, the first experiments with the blind test set-up can commence. If ADMs decisions are now chosen in a blind test in 50% of all cases or more[75], it can be considered to begin using ADMs decisions as final decisions when they are selected by the blind human evaluator as better or as good as the manual decision of the same case.

Phase four: System implementation

When ADMs decisions have been selected as the best and used as real decisions for some time, the institution can begin using a cooperative ecosystem as a fully automated decision system. In this last phase automated decisions will be sent out to citizens without any last checks. This is not without risk and therefore should be done only partly..[76] At phase four and thereafter, it is crucial, we argue, that the institution continues to operate three different decision-procedure tracks. This is to ensure that the quality of automated decisions do not deteriorate to a level below that which humans can perform. The three track system thereby serves as a model to preserve confidence in the quality of the automated decisions. The overall decision architecture can be illustrated in the following way:

We envisage that the results of the "best in test" should continuously regulate the relationship between the number of cases that are dealt with algorithmically and the number of cases that are dealt with manually. The more algorithmic drafts are preferred from or are as seen as indistinguishable from human drafts, the more cases should be dealt with algorithmically.

A key question remains how cases are distributed between the three tracks. In our first proposed model, the case load in an administrative field that is supported by ADMs is randomly split in two loads, such that one load (e.g. 80%) is fed to the algorithm for drafting and another load (e.g. 20%) is fed to a human case worker, also for drafting (this scenario could be imagined under phase 3). Drafts are subsequently sent to a head of office, who finalizes and signs off on the decisions. All final decisions are subsequently pooled and used to regularly update the algorithm used. By having human administrators interact with algorithmic drafting in this way, and feeding decisions back into the machine-learning process, the algorithm will be kept fresh with new original decisions, a percentage of which will be written by humans from scratch.

The effect of splitting the case load and leaving one part to pass through a purely manual track is that the sensitivity to the broader contextualization is fed back into the algorithm and hence allows a development in the case law that could otherwise not happen. Although human decision-making is also built from routine and former practice – that, after all is the *raison d'être* of bureaucratization – by singling out a part of the case load to be manually handled and making the human caseworkers aware of the overall working of the system, could well heighten their attention to their role in assuring that decisions are up to present day conditions. To replace human case workers with algorithms as described above will be legally compliant as long as the algorithmic drafts are of equal (high) quality as compared to the human made drafts which the model itself is designed to ensure[77]. In this way it is outcome centric while remaining human centric.

75 This percentage is only a place holder until an empirical test can be carried out to determine an acceptable limit.

76 It must also be recalled that ADM can only be put to use by fulfilling the requirements in GDPR art. 22. In the context of public authorities making decisions under the law, this means that it requires a specific legal basis in domestic law. Citizens who are made subject to ADM should also be informed of this fact.

77 Again we also emphasize that automation must respect the limits set out in art. 22. See for the most recent analysis of the requirements following from article 22, the Advocate General's Opinion in Case C-634/21 | SCHUFA Holding and Others (Scoring) and in Joint Cases C-26/22 and C-64/22 SCHUFA Holding and Others (Discharge from remaining debts) available at: https://curia.europa.eu/jcms/upload/docs/application/pdf/2023-03/cp230049en.pdf even if this case is concerned with private finance, the considerations furthered by AG Pikamäe is also relevant in the context of public administration. Note that the AG opinion is not binding for the court and that the CJEU has still not ruled in the case.

Such a cooperative intelligence approach, we think, will be optimal for providing working conditions in which ADMs, in the long-term perspective, can grow into a means for assuring more efficiency and higher quality all-round in the decision-making system. When we recall that the human-only alternative to using ADMs in public administration is nowhere near flawless and that fully automated AI systems often undermine the rule of law, using a hybrid approach that better enables continuous quality checks by human experts should help allay the fears of ADMs biggest pitfalls and make it possible to introduce ADMs into public administration without undermining trust. In an optimistic scenario, such a hybrid system may even enhance trust in the institution. In line with our analysis of trust as a matter of moral reciprocity in the sense of the trustors confidence that the trustee will act responsibly and reasonably *vis a vis* the trustor, institutional trust meaning confidence in the institution's reliable functioning can be fostered by continually, and in a conscious way adding new context into the system via the manual decision-making track. This will allow for transparency by the continuous "best in test" model. The cooperative intelligence model can add ADM's well known qualities to decision-making practices and simultaneously protect against ADMs' worst implementations.